AD_______________

Award Number: DAMD17-03-1-0046

TITLE: Sequence Variants in Estrogen Receptors and Risk for Prostate Cancer

PRINCIPAL INVESTIGATOR: Bao-Li Chang, Ph.D.

CONTRACTING ORGANIZATION: Wake Forest University Health Sciences
Winston-Salem, North Carolina 27157-0001

REPORT DATE: March 2004

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
March 2004

3. REPORT TYPE AND DATES COVERED
Annual Summary (1 Mar 2003 - 28 Feb 2004)

4. TITLE AND SUBTITLE
Sequence Variants in Estrogen Receptors and Risk for Prostate Cancer

5. FUNDING NUMBERS
DAMD17-03-1-0046

6. AUTHOR(S)
Bao-Li Chang, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Wake Forest University Health Sciences
Winston-Salem, North Carolina 27157-0001

E-Mail: bchang@wfubmc.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

This post doctoral research training aims to prepare the trainee for further prostate cancer research through a genetic epidemiology study of the role of estrogen receptors (ERs) in prostate cancer etiology. Two goals were proposed in this postdoctoral training project: 1) to further the trainee's ability to be an efficient researcher of prostate cancer genetics by enhancing the trainee's ability to use bioinformatics tools for genetic data analyses. 2) to comprehensively evaluate multiple genetic variants in ER genes using both family-based and case-control study designs, thus providing the trainee with real-world experience in genetic and bioinformatics analyses. Consistent with those goals, major accomplishments in the first half of the training period include: 1) attending statistic and bioinformatics courses, conferences, and workshops. 2) complete sequencing of all 14 exons, and the ~2kb promoter region of the ESR2 gene in 96 subjects for the identification of mutations and sequence variants. 3) application of the statistics and bioinformatics skills gained in various courses and workshops for several similar gentic epidemiological studies. The screening of sequence variants in ESR1, as well as the genotyping of both ESR1 and ESR2, are both ongoing. The analyses and results are expected during the rest of the training period.

14. SUBJECT TERMS
No Subject Terms Provided.

15. NUMBER OF PAGES
28

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
Unlimited

Table of Contents

Introduction

This postdoctoral research training aims to prepare the trainee for further prostate cancer research through a genetic epidemiology study of the role of estrogen receptors (ERs) in prostate cancer etiology. The importance of estrogens in maintaining the normal physiological status of the prostate, and potentially in the process of prostate carcinogenesis, has been clearly demonstrated in many studies. Because estrogen signaling is largely mediated by estrogen receptors (ER) alpha and beta (ESR1 and ESR2), our hypothesis is that genomic sequence variants in ER genes may affect individual susceptibility to prostate cancer. Two goals were proposed in this postdoctoral training project: 1) to further the trainee's ability to be an efficient researcher of prostate cancer genetics by enhancing the trainee's ability to use bioinformatics tools for genetic data analyses, through a formal mentorship as well as various courses and workshops. 2) to comprehensively evaluate multiple genetic variants in ER genes using both family-based and case-control study designs, thus providing the trainee with real-world experience in genetic and bioinformatic analyses. The second goal not only provides real-world experience for the trainee in the performance of genetic and bioinformatic analyses, but will also advance our knowledge of the role of the estrogen pathway in prostate cancer etiology.

Body

This postdoctoral research training grant aims to prepare the trainee for further prostate cancer research through a genetic epidemiology study of the role of estrogen receptors (ERs) in prostate cancer etiology. Major accomplishments in the first half of the training period include: 1) attending statistics and bioinformatics courses, conferences, and workshops. 2) complete sequencing of all 14 exons, and the ~2kb promoter region of the ESR2 gene in 96 subjects for the identification of mutations and sequence variants 3) application of the statistics and bioinformatics skills gained in various courses and workshops for several similar genetic epidemiological studies. The progress will be detailed in the following paragraphs.

Training in genetic data analyses and bioinformatics:

1. Statistics coursework and workshop:

   a. "Genetic Analysis Workshop 13 (GAW13)": Trainee joined the team led by several genetic epidemiologists and statisticians, including Drs. Deborah Meyers, Jianfeng Xu, Carl Langefield, and Leslie Lange, to analyze the dataset distributed by GAW in year 2002. GAWs are a collaborative effort among genetic epidemiologists to evaluate and compare statistical genetic methods. Sets of real or computer-simulated data are distributed to investigators worldwide, and the results of analyses are discussed and compared at meetings. We used a multivariate principal components method in an attempt to identify clusters of variables that may be controlled by a common gene or genes (pleiotropy) in the distributed dataset. These results were presented in the Genetic Analysis Workshop 13: Analysis of Longitudinal Family Data for Complex Diseases and Related Risk Factors, and was also published in NMC Genetics (Appendix A).

   b. The trainee attended an applied statistical methods course, titled "logistic regression", that was offered in the department of epidemiology at WFUSM (Wake Forest University School of Medicine). For this course, Dr. Carl Langefield served as Dr. Chang's mentor.

   c. The trainee also attended two SAS programming courses offered by the SAS Institute Inc. in Cary, NC. SAS is a statistics program used by most epidemiologists as a statistic analysis tool and database management system. These two courses have improved the trainee's SAS programming skills in genetic data management and analyses.

2. Bioinformatics coursework:

   a. The trainee attended "Genome Sequence Analysis: Theory and Practice" at Jackson Laboratory, in Bar Harbor, Maine. This course focused on the process of analyzing genomic sequences to identify biologically significant features using computational and comparative approaches. In addition, it also included instruction on predicting and annotating the function of genes and gene products.

b. The trainee attended a course titled, "Computational Analysis in Molecular Biology", at WFUSM. This course focuses on utilizing computer software in the analysis of gene sequences and molecular modeling.

This bioinformatics training, from both the workshop and the course, enabled the trainee to utilize various bioinformatics tools in computational analyses of the potential functional impact of the sequence variants in several candidate genes, as detailed in later paragraphs of this progress report.

3. Attended annual meetings of the American Society of Cancer Research, and the American Society of Human Genetics.

Identification of mutations and sequence variants in the ESR2 gene by sequencing all 14 exons, ~2 kb promoter region of ESR2 in 96 hereditary prostate cancer (HPC) probands.

The genomic structure of ESR2 is far more complex than what was originally described. With more and more alternatively splicing variants being found, the number of exons for ESR2 grew from 8 (referring to the cDNA sequence that encodes the full-length 530 amino acid ESR2 protein in most ESR2 studies) to 14 (NCBI annotation, build 34). A total of 16 ESR2 isoforms with alternatively spliced exons were cloned and characterized. These isoforms were believed to have different ligand and DNA binding affinity, transactivation capability, and even different subcellular localizations. Many of these isoforms and exon-skipping transcripts were found to be expressed in prostate tissues or prostate cancer cell lines. To fully characterize the genetic variants in ESR2, we screened all exons, exon-intron junctions, and ~ 2kb of the promoter region in 96 HPC probands by sequencing. In summary, there were more than 45 sequence variants identified in these subjects. Although no missense mutations were identified, some synonymous changes were implicated as having a functional impact on expression levels. In addition, more than 12 sequence variants were identified in the promoter region alone. One of the 40 bp insertion/deletion polymorphisms may increase/decrease the number of SP1 transcription factor binding elements. A subset of these 45 sequence variants will be selected, mainly by their potential functional effects as suggested by bioinfomatic analysis, for further genotyping efforts in the HPC families, as well as in the case control population.

The ESR2 sequence information was also applied to a breast cancer study recently published (appendix B). The analysis of six common sequence variants in 1134 cases and 1235 controls in the Shanghai Breast Cancer Study provided evidence for positive associations between breast cancer risk and two SNPs [C(14206)T and C(33390)G] among postmenopausal women. Evidence of a stronger association was found for SNP [C(33390)G] among women with a long duration (≥34 years) of menstruation (OR = 2.37, 95%CI = 1.18 – 4.77). A potential synergistic effect between SNP [C(33390)G] and several steroid sex hormones was observed, and a 3 to 4-fold elevated risk of breast cancer was found among women with a CG or GG genotype in SNP [C(33390)G] who also had a high level of steroid sex hormone or a low level of sex hormone binding globulin.

Application of the statistics and bioinformatics skills gained in various courses and workshops for several similar genetic epidemiological studies.

The trainee has also applied her new statistics and bioinformatics skills in several similar genetic epidemiological studies:

Genomic structure of *RNASEL*

RNase L is a single-strand-specific endonuclease that functions as the terminal enzyme component of the interferon-inducible 2-5A synthetase/RNAase L antiviral pathway. Recently, the gene encoding RNase L, *RNASEL*, was identified as a strong candidate for the prostate cancer susceptibility gene at the HPC1 locus. However, neither the public databases nor published studies have fully described and annotated the complete genomic structure of the *RNASEL* gene. From a computational comparison of all protein sequences in the databases to RNASEL, a "*RNASEL*-unique" sequence was chosen for our northern probe. With this approach, we were able to avoid cross-hybridization to homologous proteins and identify a single species of transcript at ~4.4kb. In addition, the highest expression was found to be in the prostate, followed by the testis and small intestine. The remaining tissues have very low RNASEL expression levels. To determine the total number of exons, as well as to identify the 5'- and 3'- end of the RNASEL transcript, we carried out RT-PCR, followed by 5'- and 3'-RACE. A 7-exon model, including an untranslated 1st exon which is not annotated by NCBI, was suggested both by our experimental data and by a computational annotation based on an EST database search. Interestingly, comparative genomic analyses between human *RNASEL* and mouse *Rnasel* loci identified several homologous regions in the proximal promoter and exon 1, thus indicating common transcription regulatory mechanisms between the two species (appendix C).

TLR4 isoforms and 3'untranslated region structure

There is growing evidence that chronic inflammation may also play a role in the development of prostate cancer. TLR4 is a central player in the signaling pathways that control the innate immune response and is an important candidate inflammatory gene. We performed a systematic genetic analysis of TLR4 sequence variants by evaluating eight SNPs that span the entire gene, among 1383 newly diagnosed prostate cancer patients and 780 age- and residence-matched controls in Sweden. We found an association between a sequence variant (11380G/C) in the 3'UTR of the TLR4 gene and prostate cancer risk (Appendix D). Although the location of SNP 11380G/C is controversial, it is likely within the 3'UTR of isoform A. There are two reported human TLR4 cDNA sequences, and both are deposited in GenBank (U93091 and U88880). The cDNA clone U93091 (4874 bp) corresponds to isoform A and contains exons 1, 3, and 4. However, it has an extra 1096 bp in the 3'UTR region, compared to the NCBI RefSeq for isoform A. The cDNA clone U88880 (3811 bp) contains all four exons, with the same, shorter, exon 4 sequence that is annotated in the NCBI database. The cDNA clone U88880 may use a downstream translation start site in exon 3 and encode a shorter 799 a.a. isoform C. Apparently, the current NCBI database used clone U88880 as a template and did not use the information from clone U93091 for annotation of the 3'UTR. To preliminarily define the 3'UTR region using computational methods, we used either 2kb or 5kb of additional genomic sequence 5' as well as 3' beyond the end of the annotated genomic sequence, in a BLAST search against the human EST database. Eight EST clones from various sources were aligned to the 3' region that followed the NCBI annotated 4th exon and overlapped with the 4th exon. These EST sequences only have significant similarity to the TLR4 genomic sequence in the whole genome when used as queries in BLAST. The evidence from EST clones extended for an additional 568 bp into the 3' region of TLR4, and the SNP 11380G/C was mapped to this 3' region. This analysis provided further evidence that the longer 3'UTR of TLR4 is real. Interestingly, seven additional EST clones were uniquely aligned to the genomic region extending approximately 1123 to 1460 bp beyond the end of the annotated 4th exon. This EST evidence extends the 3' UTR region of TLR4 an additional 27 to 364 bp, compared to the longest 3'UTR in clone U93091. Collectively, this indicates that either U93091 does not represent a full length clone, or that there is another isoform which differs from U93091 by having an extra 3' UTR region. Further experimental data is needed to support these conclusions drawn from computational methods.

Extronic splice elements in *ESR2* gene

For genetic variants in the coding region, attention has been primarily focused on the nonsense and missense changes that alter protein sequences and functions. However, there is increasing evidence that many human disease genes harbor exonic mutations that affect pre-mRNA splicing. Accumulating evidence shows that nucleotide changes in the exons may inactivate genes by inducing the splicing machinery to skip the mutant exons, and this has even been observed for some nonsense, nonsynonymous, or even silent synonymous mutations. Similarly, SNPs in coding-regions might cause phenotypic variability by influencing splicing accuracy or efficiency. When we studied G(25652)A in exon 6 and C(33390)G in exon 7 of *ESR2* in a breast cancer population, we explored the possibility that these two synonymous changes may affect the splicing of the ESR2 pre-mRNA by searching the exonic splicing enhancer (ESE) motifs. Both SNPs were found to be in ESE motifs. G(25652)A was found to be in the binding site for splice factor SRp55, and C(33390)G was in the binding site for splice factor SC35. While the nucleotide change for G(25652)A is conserved, the nucleotide change for C(33390)G may have an influence on the splice efficiency of exon 7. The C allele of C(33390)G is found preferentially in the SC35 enhancer element at this particular position (has above background frequency in SC35 consensus sequence), and the G allele of C(33390)G has below background frequency in this position and may have an adverse effect on splicing efficiency. Many splice variants have been identified in ESR2, and differential expression of these splice variants during breast carcinogenesis has also been reported. The differential ligand specificity among splice variants has been explored because it has been hypothesized to affect the efficacy of various selective estrogen receptor modulators (SERMs). It is likely that SNP C(33390)G, which may impact the splicing efficiency of exon 7, may alter the expression pattern of ESR2 splice variants and therefore the risk of breast cancer (appendix B).

Key Research Accomplishments

The major accomplishments in the first half of the training period include:
1) Attending statistic and bioinformatic courseworks, conferences, and workshops.
2) Complete sequencing all 14 exons, and ~2 kb promoter region of *ESR2* gene in 96 hereditary prostate cancer probands for identification of mutations and sequence variants.
3) Utilizing and incorporating the statistic and bioinformatic skills gained in various courseworks and workshops in several similar genetic epidemiological studies.

Reportable outcomes

Several manuscripts and meeting abstract published during the first half of this training period:

Bensen JT, Lange LA, Langefeld CD, Chang B, Bleecker ER, Meyers DA, Xu J. Exploring Pleiotropy using Principal Components. BioMed Central Genetics 2003; 4:S53.

Zheng SL, Zheng W, Chang BL, Shu XO, Cai Q, Yu H, Dai Q, Xu J, Gao YT. Joint effect of estrogen receptor beta sequence variants and endogenous estrogen exposure on breast cancer risk in Chinese women. Cancer Res. 2003;63:7624-9.

Chang BL, Tolin A, Segade F, Silverman R, Isaacs WB, Xu J. Genomic structure and the expression patterns of RNASEL. AACR proceeding 2004.

Zheng SL, Augustsson-Bälter K, Chang BL, Hedelin M, Li L, Adami HO, Bensen J, Li G, Johnasson JE, Tolin A, Turner AR, Adams TS, Meyers DA, Isaacs W, Xu J, Grönberg H. Sequence variants of TLR4 are associated with prostate cancer risk: results from the CAPS study. Cancer Res 2004 (In press)

Conclusions

In summary, the major accomplishments in the first half of the training period include: 1) attending statistics and bioinformatics courses, conferences, and workshops. 2) complete sequencing of all 14 exons, and the ~2kb promoter region of the *ESR2* gene in 96 subjects for the identification of mutations and sequence variants. 3) application of the statistics and bioinformatics skills gained in various courses and workshops for several similar genetic epidemiological studies. The screening of sequence variants in ESR1, as well as the genotyping of both ESR1 and ESR2, are both ongoing. The analyses and results are expected during the rest of the training period.

Appendix A

BMC Genetics

BioMed Central

Proceedings

Open Access

Exploring pleiotropy using principal components

Jeannette T Bensen*†[1], Leslie A Lange†[2], Carl D Langefeld[2], Bao-Li Chang[1], Eugene R Bleecker[1], Deborah A Meyers[1] and Jianfeng Xu[1,2]

Address: [1]Center for Human Genomics, Wake Forest University School of Medicine, Winston-Salem, North Carolina, USA and [2]Department of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, North Carolina, USA

Email: Jeannette T Bensen* - jbensen@wfubmc.edu; Leslie A Lange - llange@wfubmc.edu; Carl D Langefeld - clangef@wfubmc.edu; Bao-Li Chang - bchang@wfubmc.edu; Eugene R Bleecker - ebleeck@wfubmc.ed; Deborah A Meyers - dmeyers@wfubmc.edu; Jianfeng Xu - jxu@wfubmc.edu

* Corresponding author †Equal contributors

from Genetic Analysis Workshop 13: Analysis of Longitudinal Family Data for Complex Diseases and Related Risk Factors
New Orleans Marriott Hotel, New Orleans, LA, USA, November 11–14, 2002

Published: 31 December 2003

BMC Genetics 2003, 4(Suppl 1):S53

This article is available from: http://www.biomedcentral.com/1471-2156/4/s1/S53

Abstract

A standard multivariate principal components (PCs) method was utilized to identify clusters of variables that may be controlled by a common gene or genes (pleiotropy). Heritability estimates were obtained and linkage analyses performed on six individual traits (total cholesterol (Chol), high and low density lipoproteins, triglycerides (TG), body mass index (BMI), and systolic blood pressure (SBP)) and on each PC to compare our ability to identify major gene effects. Using the simulated data from Genetic Analysis Workshop 13 (Cohort 1 and 2 data for year 11), the quantitative traits were first adjusted for age, sex, and smoking (cigarettes per day). Adjusted variables were standardized and PCs calculated followed by orthogonal transformation (varimax rotation). Rotated PCs were then subjected to heritability and quantitative multipoint linkage analysis. The first three PCs explained 73% of the total phenotypic variance. Heritability estimates were above 0.60 for all three PCs. We performed linkage analyses on the PCs as well as the individual traits. The majority of pleiotropic and trait-specific genes were not identified. Standard PCs analysis methods did not facilitate the identification of pleiotropic genes affecting the six traits examined in the simulated data set. In addition, genes contributing 20% of the variance in traits with over 0.60 heritability estimates could not be identified in this simulated data set using traditional quantitative trait linkage analyses. Lack of identification of pleiotropic and trait-specific genes in some cases may reflect their low contribution to the traits/PCs examined or more importantly, characteristics of the sample group analyzed, and not simply a failure of the PC approach itself.

Background

Principal component analyses often provide valuable information that allows data reduction and reveals relationships between variables that were not previously suspected. As we begin to better understand the scope of gene effects, we find that single genes often contribute to multiple phenotypes (pleiotropy). Therefore, when mapping genes for complex disorders, it can be helpful to identify groups of variables or phenotypes (principal components) that may be controlled by a single gene. Arya et al. [1] demonstrated the practical application of principal component analysis by evaluating eight insulin resistance syndrome-related phenotypes in 27 nondiabetic Mexican-American extended families [1]. In their analyses, they identified three principal components factors and following multipoint variance components linkage analyses, their adiposity-insulin factor showed linkage at two different regions on chromosome 6q with LOD scores > 4.1. This observation was consistent with their previous finding of a major susceptibility locus for insulin resistance on chromosome 6q, which has been shown to have strong pleiotropic effects on other insulin resistance syndrome-related phenotypes such as body mass index (BMI) and leptin levels [1,2]. To examine this type of pleiotropic gene effect seen in the Arya study, we chose to evaluate the use of standard principal components (PC) methods to capture this effect in the Genetic Analysis Workshop (GAW13) simulated data set. Our first objective was to assess whether the traits grouped together in one of the PCs in our data analysis actually correspond to traits that share common gene effects in the underlying GAW13 simulated genetic model, our second objective was to identify the heritability of these PCs, and our third objective was to identify major pleiotropic genes through linkage analysis.

Methods

All analyses were performed on the simulated data without missing observations. Replicate data set 57 was randomly selected and analysis was limited to the year 11 time point. Year 11 was selected because this was the first year in which Cohorts 1 and 2 both had data collected. Observations with triglyceride values greater than 400 (n = 27) were excluded in order to obtain valid low density lipoprotein (LDL) calculations. In addition, several (n = 8 > 4 SD) observations were excluded because they were judged to be highly influential in the PC analysis.

PC analysis was conducted on six quantitative traits (QTs): total cholesterol (Chol), triglycerides (TG), high density lipoprotein (HDL), LDL, systolic blood pressure (SBP), and body mass index (BMI). LDL was calculated using the Friedewald's equation [3]: (Total Chol - HDL) - TG/5, where TG = 400. BMI was calculated as (weight (lb) / (height (in))^2] * 703. Three of the QTs (Chol, TG, and SBP) were log-transformed in order to better conform to a normal distribution. Each QT was then regressed on sex, age, and cigarettes per day using linear regression modeling, and residuals were obtained. The residuals for each QT were then standardized. PCs were calculated from the correlation matrix of the standardized residuals corresponding to the six QTs using standard methods, in which all individuals are assumed to be independent. PC analysis was performed using PROC FACTOR in the SAS statistical software package (version 8.2, Cary, NC), with PC extraction and varimax rotation (Table 1). Results from this analysis were used to create PCs consisting of linear combinations of individual QT residuals.

Heritability estimation and quantitative multipoint linkage analysis were performed on the PCs and on the residuals for individual QTs using variance-component methodology, as implemented in the Sequential Oligogenic Linkage Analysis Routines (SOLAR) [4]. Genotype data provided from all individuals were used to generate multipoint identity-by-descent (IBD) estimates throughout the genome. Phenotypic traits examined included the PCs and the raw QTs. No additional covariate adjustment was made at this stage. All analyses were performed a second time, with additional adjustment for cohort effect (using an indicator variable) when residuals were obtained. This was done in order to examine whether cohort had an effect after adjusting for age.

We did not consult the GAW13 simulated data set answers prior to either the interpretation of the PCs or performing linkage analysis. Verification of genes modeled in the simulated data set at baseline (not those influencing longitudinal data) were considered verified if linkage analysis identified a marker with a peak LOD score (LOD > 1.0) within 20 cM of the gender-averaged chromosomal location for a simulated trait gene. While there is little consensus regarding the most appropriate LOD score threshold for complex disease, similar to other studies of complex disease reporting LODs less than 2.0, we considered LOD scores greater than 1.0 as suggestive evidence of linkage [5,6].

Results

At year 11 we had complete data on 989 individuals (316 families) from Cohort 1, mean age 59.9 years, and 1511 individuals (330 families) in Cohort 2, mean age 53.4. Variable means for the QTs and confounders were comparable between cohorts, except for SBP, TGs, and cigarettes per day, where mean SBP and TG were higher in Cohort 1 than 2 (SBP: 137 vs. 130 and TG: 146 vs. 136, respectively) and mean cigarettes per day were lower in Cohort 1 than 2 (4 vs. 6, respectively). After adjustment for age, sex, and cigarettes per day, cohort was a statistically significant predictor of only one of the QTs: SBP. The additional adjustment for cohort produced results (PCs and linkage) that were similar to those reported and did not change any of our conclusions.

The first three principal components identified in this analysis contributed to 73% of the overall phenotypic variance among the six QTs (Table 2). Heritability estimates (polygenic) for individual QTs and the three primary principal components were all statistically significant ($p <$ 0.0001), ranging from 0.60 for LDL to 0.79 for BMI (Table

Table 1: Principal component trait loading values (rotated values).

Trait	PC1	PC2	PC3
Log Chol	0.96	0.03	0.08
HDL	-0.17	-0.73	0.21
LDL	0.99	0.01	-0.03
Log TG	-0.11	0.79	0.16
BMI	-0.04	0.44	0.54
Log SBP	0.07	-0.15	0.83

Major determinants of the PC were considered traits with loading values = 0.30.

Table 2: Variance and heritability estimates for individual traits and principal components.

	Trait	Mean (SD)	H2r[A] (SE)
1	Log Chol	5.32 (0.17)	0.63 (0.04)
2	HDL	50.22 (11.64)	0.71 (0.03)
3	LDL	128.89 (37.90)	0.60 (0.04)
4	Log TG	4.83 (0.48)	0.62 (0.04)
5	BMI	26.74 (4.79)	0.79 (0.03)
6	Log SBP	4.88 (0.12)	0.75 (0.03)
PC 1	1/2 (Log Chol + LDL)		0.62 (0.04)
PC 2	1/2 BMI + (Log TG - HDL)		0.80 (0.03)
PC 3	Log SBP + 2/3 BMI		0.74 (0.03)

[A] H2r represents the polygenic contribution and H2q1 the contribution of major gene (H2r + H2q1 = overall heritability).

Table 3: Genome-wide linkage results for principal components.

Component	Peak	Maximum LOD	Chromosome	Position (cM)	Marker
PC 1	---	No LOD > 1.0			
PC 2	---	No LOD > 1.0			
PC 3	Peak 1	1.18	3	132	False +
	Peak 2	1.07	7	137	b10 @ 124 (height)
	Peak 3	1.16	15	20	False +

2). Standard errors for the heritabilities for all QTs and PCs were typically between 0.03 and 0.04

For PCs, linkage analysis only yielded LOD scores greater than 1.0 but less than 2.0 for PC3 (SBP + 2/3 BMI). Two of the three LODs in this range were false-positive results according to our criteria, while the third LOD identified a minor gene (b10) contributing 1% of trait variation for height (Table 3).

For individual traits, no LOD scores > 1.0 were observed for log Chol, HDL, LDL, or log SBP (Table 4). Log TG yielded two LOD scores between 1.0 and 2.0, both of which were false-positive findings, while BMI produced 31 LOD scores > 1.0, with 4 scores > 2.0. When considering the LODs between 1.0 and 2.0 for BMI, 26 of 27 (96%) were false-positive results, while 1 LOD score identified a gene for height, a component of the BMI quantitative trait. Of the 4 LOD scores greater than 2 for BMI, 2 were false positive, 1 was essentially unrelated to the BMI trait identifying genes for cholesterol and HDL, while only the highest LOD (5.4) identified a gene contributing 40% to trait variance for weight.

Table 5 indicates the linkage results within 20 cM of the two pleiotropic genes, b12 and b13, that contribute the largest proportion to the phenotypic variance of both

Table 4: Genome-wide linkage results for individual traits.

Trait	Peak	Maximum LOD	Chromosome	Position (cM)	Marker
Log Chol	---	No LOD > 1.0			
HDL	---	No LOD > 1.0			
LDL	---	No LOD > 1.0			
Log TG	Peak 1	1.55	12	166	False +
	Peak 2	1.02	19	68	False +
Log SBP		No LOD > 1.0			
BMI	Peak 8	3.11	4	35	False +
	Peak 19	2.74	11	50	b30 @66 (Chol) b21 @45 (HDL)
	Peak 23	5.40	13	55	b11 @70 (Weight)
	Peak 26	2.21	15	15	False +

TG = triglycerides, Chol = Cholesterol, SBP = systemic blood pressure, BMI = body mass index

Table 5: Unblinded major pleiotropic genes influencing TG and HDL-linkage results.

	PC 1 1/2 (Log Chol + LDL)			PC 2 PC 2 1/2 BMI + (Log TG - HDL)			PC 3 PC 3 Log SBP + 2/3 BMI		
Gene	**Max LOD[A]**	**H2r[B]**	**H2q1[B]**	**Max LOD[A]**	**H2r[B]**	**H2q1[B]**	**Max LOD[A]**	**H2r[B]**	**H2q1[B]**
G(b12)[C]	0.00	0.62	0.00	0.00	0.80	0.00	0.42	0.71	0.05
G(b13)[D]	0.00	0.62	0.00	0.00	0.80	0.00	0.24	0.72	0.03

[A] MaxLOD, the maximum LOD score within approximately 20 cM of the gene.
[B] H2r represents the polygenic contribution and H2q1 the contribution of major gene (H2r + H2q1 = overall heritability).
[C] G(b12) is located on chromosome 9 at 11 cM (MaxLOD range: 0 cM-35 cM).
[D] G(b13) is located on chromosome 9 at 83 cM (MaxLOD range: 65 cM-105 cM).

HDL and TG. No elevated LOD score > 1.0 was identified for either PC1, PC2, or PC3.

Discussion

Pleiotropic effects are a common phenomenon in reported studies of complex disease. Methods are needed to identify pleiotropic genes that may contribute differing amounts to the variances of multiple phenotypes. To this end, we chose to evaluate our ability to identify such genes by PC analysis, followed by heritability estimates and linkage analysis.

While our analysis was somewhat limited in terms of the number of variables available in the *complete* data set, PC analysis of the six variables identified three primary PCs explaining 79% of the phenotypic variance. Covariates (age, gender, and smoking) were adjusted *prior* to PC analysis, consistent with the strategy used by Moser et al., although concerns about the effect of these adjustments on PC and heritability estimates arose [7]. We therefore performed covariate adjustments *before* and *after* PC analysis [data not shown] and found no significant differences in PCs, loading, or heritability. Overall, the PC analysis, in particular PC2, reflected the pleiotropic genes (HDL and TG) modeled in the simulated data.

Heritability estimates were statistically significant for each of the three major PCs, as were those for the traits evaluated individually. Each PC heritability estimate was consistent in magnitude with the trait heritabilities comprising the PC. PC2, which reflected the simulated model best with respect to shared gene effects, had a heritability estimate slightly higher than the two individual variables (HDL and TG) in the PC and closer to that for BMI alone. This higher heritability estimate for PC2 may reflect the accuracy with which PC identifies/groups variables with common genetic influence or it may reflect the significant influence of BMI on this PC.

Several factors that may have contributed to limited power in both our individual trait and PC linkage analyses include sample size and composition (single replicate), pedigree structure, and the number and size of genetic effects. One of the challenges facing linkage mapping for complex disease traits is adequate sample size. Risch and Merikangas state that the power of linkage for complex disease is limited to the detection of only the strongest loci unless thousands of small families are utilized [8]. In this report a total of 646 families were analyzed and thus may not have provided ample power for the detection of genes contributing modestly to trait variance. The analysis of a single replicate in the GAW13 simulated data set may also have hindered our ability to detect meaningful linkage.

Studies have shown the PC approach may improve the power to identify genes with pleiotropic effects involved in complex disease [1,9,10]. While PC heritability estimates were encouraging, we were unable to identify pleiotropic genes. One very plausible explanation may be that rather than a single gene with a major effect, the high heritability reflected many genes with small effects. While it has been shown that the PC approach has greater power to detect major pleiotropic genes [10], the power to detect genes with small effects is likely to be limited. In addition, our investigation was highly dependent on the extent of pleiotropy modeled in the simulated data set as well as our selection of variables for analysis. HDL, TG, and glucose were modeled as pleiotropic traits; however our investigation only considered HDL and TG (major components of our PC2). Ideally, PC2 would have identified at least the b12 gene contributing 20% and 10% to the variance of HDL and TG, respectively. Several investigators have demonstrated increased power and precision in identifying genetic effects when using multivariate approaches for correlated traits [11,12]. However, in a recent commentary, Meigs points out that the results of such analyses can be influenced by both the number and nature of variables included in the model [13]. The lack of our ability to identify the b12 gene in this simulated data set may have been due to the omission of glucose from our model or may reflect the difficulty our method has in identifying complex trait genes. Finally, while we utilized the standard PC method and adjusted for covariates prior to linkage analysis to maximize power, we may have missed potentially important genetic effects by focusing first on the PCs that explained the majority of phenotypic variation.

In summary, PC analysis has been demonstrated in reported studies of complex disease to localize regions of the human genome likely to contain pleiotropic genes [1], but may be influenced by factors such as the number and effect size of pleiotropic genes involved as well as complex trait variables available for inclusion in the PC analysis. Further studies are needed to assess the utility of the PC approach in complex disease.

Acknowledgments

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI.

References

1. Arya R, Blangero J, Williams K, Almasy L, Dyer TD, Leach RJ, O'Connell P, Stern MP, Duggirala R: **Factors of insulin resistance syndrome-related phenotypes are linked to genetic locations on chromosomes 6 and 7 in nondiabetic Mexican-Americans.** *Diabetes* 2002, **51:**841-847.
2. Duggirala R, Blangero J, Almasy L, Arya R, Dyer TD, Williams K, Leach RJ, O'Connell P, Stern MP: **A major locus for fasting insulin concentrations and insulin resistance on chromosome 6q with strong pleiotropic effects on obesity-related phenotypes in non-diabetic Mexican-Americans.** *Am J Hum Genet* 2001, **68:**1149-1164.
3. Friedewald WT, Levy RI, Fredrickson DS: **Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge.** *Clin Chem* 1972, **18:**499-502.
4. Almasy L, Blangero J: **Multipoint quantitative trait linkage analysis in general pedigrees.** *Am J Hum Genet* 1998, **62:**1198-1211.
5. Silverman EK, Palmer LJ, Mosley JD, Barth M, Senter JM, Brown A, Drazen JM, Kwiatkowski DJ, Chapman HA, Campbell EJ, Province MA, Rao DC, Reilly JJ, Ginns LC, Speizer FE, Weiss ST: **Genomewide linkage analysis of quantitative spirometric phenotypes in severe early-onset chronic obstructive pulmonary disease.** *Am J Hum Genet* 2002, **70:**1229-1239.
6. Angius A, Petretto E, Maestrale GB, Forabosco P, Casu G, Piras D, Fanciulli M, Falchi M, Melis PM, Palermo M, Pirastu M: **A new essential hypertension susceptibility locus on chromosome 2p24-p25, detected by genomewide search.** *Am J Hum Genet* 2002, **71:**893-905.
7. Moser KL, Jedrey CM, Conti D, Schick JH, Gray-McGuire C, Nath SK, Daley D, Olson JM: **Comparison of three methods for obtaining principal components from family data in genetic analysis of complex disease.** *Genet Epidemiol* 2002, **21(suppl 1):**S726-S731.
8. Risch N, Merikangas K: **The future of genetic studies of complex human diseases.** *Science* 1996, **273:**1516-1517.
9. Morton NE, Matsuura J, Bart R, Lew R: **Genetic epidemiology of an institutionalized cohort of mental retardates.** *Clin Genet* 1978, **13:**449-461.
10. Ott J, Rabinowitz D: **A principal-components approach based on heritability for combining phenotype information.** *Hum Hered* 1999, **49:**106-111.
11. Amos CI, de Andrade M, Shu D: **Comparison of multivariate tests for genetic linkage.** *Hum Hered* 2001, **51:**133-144.
12. Evans DM: **The power of multivariate quantitative-trait loci linkage analysis is influenced by the correlation between the variables.** *Am J Hum Genet* 2002, **70:**1599-1602.
13. Meigs JB: **Invited commentary: Insulin resistance syndrome? Syndrome X? Multiple metabolic syndrome? A syndrome at all? Factor analysis reveals patterns in the fabric of correlated metabolic risk factors.** *Am J Epidemiol* 2000, **152:**908-911.

Advances in Brief

Appendix B

Joint Effect of Estrogen Receptor β Sequence Variants and Endogenous Estrogen Exposure on Breast Cancer Risk in Chinese Women

S. Lilly Zheng,[1] Wei Zheng,[2] Bao-li Chang,[1] Xiao-Ou Shu,[2] Qiuyin Cai,[2] Herbert Yu,[3] Qi Dai,[2] Jianfeng Xu,[1] and Yu-Tang Gao[4]

[1]Center for Human Genomics, Wake Forest University School of Medicine, Winston-Salem, North Carolina; [2]Department of Medicine and Vanderbilt-Ingram Cancer Center, School of Medicine, Vanderbilt University, Nashville, Tennessee; [3]Department of Epidemiology and Public Health and Yale Cancer Center, Yale University School of Medicine, New Haven, Connecticut; and [4]Department of Epidemiology, Shanghai Cancer Institute, Shanghai, China

Abstract

Long-term estrogen exposure and family history of breast cancer are the two factors that are most consistently found to be associated with breast cancer risk. Sequence variants in genes involved in estrogen synthesis, metabolism, and signal transduction may account, in part, for this observation. Using data and DNA samples from the Shanghai Breast Cancer Study, we tested the hypothesis that sequence variants of the estrogen receptor β gene (*ESR2*) may be associated with increased risk for breast cancer, particularly among women who have a high level and long-term endogenous estrogen exposure. Direct sequencing of the *ESR2* gene among 30 Chinese women revealed eight sequence variants. Association analysis of six common sequence variants in 1134 cases and 1235 controls provided evidence for positive associations between breast cancer risk and two single nucleotide polymorphisms (SNPs), [C(14206)T and C(33390)G], among postmenopausal women. Evidence of a stronger association was found for SNP [C(33390)G] among women with a long duration (≥34 years) of menstruation (odds ratio, 2.37; 95% confidence interval, 1.18–4.77). A potential synergistic effect between SNP [C(33390)G] and several steroid sex hormones was observed, and a 3–4-fold elevated risk of breast cancer was found among women with a CG or GG genotype in SNP [C(33390)G] combined with a high level of steroid sex hormone or a low level of sex hormone binding globulin. Our results are consistent with the hypothesis of a joint effect of estrogen receptor β sequence variants and endogenous estrogen exposure on breast cancer risk.

Introduction

Breast cancer is the leading cancer among women in most parts of the world, including Shanghai, the largest industrialized city in China. The annual incidence rate for breast cancer was 27.5 per 100,000 in Shanghai during 1993–94, approximately one-third of the rate for their Caucasian counterparts in the United States (1). The etiology for this common malignancy remains largely unknown. Of the many risk factors identified for breast cancer, long-term estrogen exposure and family history of breast cancer, are the two factors most consistently reported from previous studies, including our studies in Shanghai (2). In our recent study from Shanghai, we have shown that women with breast cancer had substantially higher blood estrogen levels than controls, and a >2-fold elevated risk of breast cancer was found among postmenopausal women in the upper tertile of blood testosterone or estrone, compared with those in the low tertile of these steroid sex hormones (3).

Similar to other complex diseases, it is likely that a combination of genetic susceptibility and exposure to endogenous and exogenous factors increase the risk to breast cancer. Mutations in high penetrance cancer susceptibility genes, such as the *BRCA1* and *BRCA2* genes, confer a substantially elevated risk; however, they only account for <10% of breast cancer cases in the general population due to their low mutation frequencies. Other genes, particularly those involved in estrogen synthesis, metabolism, and signal transduction, may also be important in the etiology of breast cancer. Estrogen signaling is largely mediated by ERs[5] α and β (ESR1 and ESR2); both are members of the nuclear receptor superfamily of ligand-inducible transcription factors. Although the *ESR1* and *ESR2* genes have a high sequence similarity in their DNA- and ligand-binding domains, they have distinct transcriptional activating function-1 domains and are believed to possess different transcriptional activation properties (4). *ESR1* and *ESR2* have distinct tissue and cell expression patterns. In normal breast tissue, *ESR2* is constitutively expressed and is the predominant ER in most cells. Various *ESR2* splicing variants have been found, and some of the splicing variants have been shown to encode proteins in both normal and cancerous tissues (5, 6). In breast cancer, the presence of ESRs is correlated with higher responsiveness to hormone therapy and better prognosis. Thus far, the majority of epidemiological association studies have focused on evaluating the association between genetic polymorphisms in *ESR1* and breast cancer risk. However, the results for several commonly examined *ESR1* genetic polymorphisms have been mixed (7). We reported recently some results from a large population-based case-control study among Chinese women in Shanghai, showing that breast cancer risk was associated with the *Pvu*II polymorphism in intron 1 (8) and a GT dinucleotide repeat polymorphism in the promoter region of the *ESR1* gene (9). In the current study, we evaluated the hypothesis that certain sequence variants of the ER β gene (*ESR2*) are associated with an increased risk for breast cancer, particularly among women who have a high level and long-term of estrogen exposure. This study includes the following components: (*a*) direct sequencing of the exons, exon-intron junctions, and the promoter region to identify sequence variants of the *ESR2* gene among 30 Chinese women; (*b*) genotyping seven relatively common sequence variants among all of the cases and controls whose DNA samples were available at the time of this study; and (*c*) assessment of the association of *ESR2* sequence variants and their interaction with estrogen exposure on breast cancer risk.

Received 7/8/03; revised 8/21/03; accepted 8/26/03.

Grant support: USPHS Grants R01CA64277 and R01CA90899 from the National Cancer Institute (W. Z.); partially supported by a grant from the Department of Defense to (J. X.).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Wei Zheng, Center for Health Services Research, Vanderbilt University Medical Center, Medical center East, Suite 6000, Nashville, TN 37232-8300. E-mail: wei.zheng@vanderbilt.edu or Jianfeng Xu, Center for Human Genomics, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157. E-mail: jxu@wfubmc.edu.

[5] The abbreviations used are: ER, estrogen receptor; CI, confidence interval; UTR, untranslated region; dNTP, deoxynucleotide triphosphate; DHEA, dehydroepiandrosterone; S. sulfate; DSL, Diagnosed Systems Laboratory Inc.; LD, linkage disequilibrium; RR, relative risk; BMI, body mass index; SNP, single nucleotide polymorphisms.

Materials and Methods

Study Population. Included in the study were subjects recruited during 1996–1998 in the Shanghai Breast Cancer study, a population-based case-control study of Chinese women in Shanghai. A detailed description of the study was provided elsewhere (2, 10). Briefly, 1459 incident breast cancer patients 25–64 years of age were enrolled in the study, along with 1556 healthy control women who had a similar age distribution to the cases based on frequency match. During the study, newly diagnosed breast cancer patients were identified through a rapid case-ascertainment system, supplemented by the population-based Shanghai Cancer registry. The cases enrolled in our study represented 91% of the newly diagnosed breast cancer patients identified for our study during the study period. The controls were randomly selected from the general population in Shanghai, using resident registration information provided by the Shanghai Resident Registry, which registers all of the permanent residents in urban Shanghai. Before random selection, we first determined the number of controls needed in each age group, at 5-year intervals, based on the number of cases in the corresponding age group reported to the Shanghai Cancer Registry in recent years. Once the numbers were determined, potential controls were randomly selected using their resident registration number. After the study eligibility of the identified potential control was confirmed, and an in-person interview was scheduled and conducted by a trained interviewer. Of those who were eligible for study, 90% completed an in-person interview.

The in-person interview was done with the use of a structured questionnaire, which elicited information on demographic features, menstrual and reproductive history, use of sex steroid hormones, medical history, physical activity, alcohol and tobacco use, dietary habits, and family history of cancer. Of the 1459 cases and 1556 controls who completed the in-person interview, morning fasting blood samples were collected from 1193 cases (82%) and 1310 controls (84%). The blood samples were processed to separate plasma within 6 h of collection, and the plasma specimens were immediately stored at −70°C. DNA samples of 2369 subjects were available at the time of the study.

We measured sex steroid hormones in a subset of cases and controls. Cases included in this substudy were those whose blood samples were collected before any cancer treatment. All of the postmenopausal cases and controls and 171 premenopausal case-control pairs were included in the study. For premenopausal women, the cases and controls were matched individually on (+5 years) and their menstrual cycle, which was either within the first 10 days of the menstrual cycle, matching mainly on follicular phase, or within 3 days after the first 10 days, matching either on follicular phase or luteal phase. To control for potential between assay variability, we adjusted the batch of hormone assay in data analysis.

Genotyping Methods. The PCR products of 9 exons, exon-intron junctions, the promoter region, and 3′ UTR of *ESR2* were sequenced directly in 15 early onset breast cancer cases and 15 controls. The primers for PCR are available.[6] All of the PCR reactions were performed in a 10-μl volume consisting of 30 ng genomic DNA, 0.2 mM of each primer, 0.2 mM of each dNTP, 1.5 mM $MgCl_2$, 20 mM Tris-HCl, 50 mM KCl, and 0.5 unit of Taq polymerase (Life Technologies, Inc.). PCR cycling conditions were as follows: 94°C for 4 min; followed by 30 cycles of 94°C for 30 s, specified annealing temperature for 30 s, and 72°C for 30 s; with a final extension of 72°C for 6 min. All of the PCR products were purified using the QuickStep PCR purification kit (Edge BioSystems, Gaithersburg, MD) to remove dNTPs and excess primers. All of the sequencing reactions were performed using dye-terminator chemistry (BigDye; ABI, Foster City, CA) and then precipitated using 63 ± 5% ethanol. Samples were loaded onto an ABI 3700 DNA Analyzer after adding 8 μl of formamide. SNPs were identified using Sequencher software version 4.0.5 (Gene Codes Corporation).

SNP genotyping was performed at the Center for Human Genomics, Wake Forest University School of Medicine. All of the investigators at Wake Forest University were blinded to the case-control status. To ensure the quality of genotyping, each 96-well DNA plate was embedded with two duplicates of our study subjects, two duplicates of Centre d'Etude du Polymorphisme Humain controls (1347–02), and one water blank. All of the duplicates were later determined to be identical for all seven of the SNPs, and no genotype was observed for any of the blanks. SNPs were genotyped using the MassARRAY system (SEQUENOM, Inc., San Diego, CA). PCR reactions were performed in a total volume of 5 μl with 10 ng of genomic DNA, 2.5 mM of $MgCl_2$, 0.1 unit of HotStarTaq polymerase (Qiagen Inc., Valencia, CA), 200 μM of dNTP, and 200 nM of primers. The PCR reactions started at 95° for 15 min, followed by 45 cycles of 95° for 20 s, 50° for 30 s, and 72° for 1 min, with a final extension of 72° for 3 min. The homogenous Mass Extend (hME) reactions were performed in a total volume of 9 μl with 50 μM each deoxynucleotide triphosphate/dideoxynucleotide triphosphate (d/ddNTP), 0.063 μ/μl of Thermo Sequenase (both from SEQUENOM, Inc.), and 600 nM of extension primers. The cycling conditions were 94° for 2 min, followed by 55 cycles of 94° for 5 s, 52° for 5 s, and 72° for 5 s. After cleaning up the hME reaction products with the SpectroCLEAN, the products were transferred to a SpectroCHIP using SpectroPOINT, and then scanned through SpectroREADER. Genotyping was done using SpectroTYPER.

Measurement of Steroid Hormones and Sex Hormone-Binding Globulin (SHBG). A detailed description of the measurement of steroid hormones and SHBG was provided elsewhere (3). Plasma concentrations of testosterone, estradiol, estrone, estrone sulfate, DHEA-S, and progesterone were measured directly without extraction. Measurement of steroids and SHBG in our study was performed in a reference laboratory, at DSL (Webster, TX). Commercial RIAs from DSL were used for the measurement of steroids; an immunoradiometric assay from DSL was used for SHBG. Technicians who performed the tests did not know the source of the specimens.

Statistical Methods. Hardy-Weinberg Equilibrium tests for each sequence variant and pair-wise LD tests among all of the sequence variants were performed using the Fisher probability test statistic, as described by Weir (11). For each test, 10,000 permutations were performed, and the test statistic of each replicate was calculated. Empirical *P*s for each test were estimated as the proportion of replicates that were as probable as or less probable than the observed data, as implemented in the software package Genetic Data Analysis. Various estimates of pair-wise LD such as Lewontin's D' and the correlation coefficient were calculated using the computer software SAS/Genetics.

Tests for association between sequence variants and breast cancer were performed by comparing the allele frequencies between cases and controls in a χ^2 test with 1 degree of freedom using the computer software SAS/Genetics. Risk genotypes were arbitrarily defined based on the alleles that were more common in cases than in controls and assuming a dominant model; *i.e.*, homozygous and heterozygous variant alleles as risk genotypes, and homozygous nonrisk allele as a reference genotype. The choice of dominant models was primarily because of the few homozygous risk allele carriers in this population. Estimates of RR were calculated and adjusted for potential confounders.

Association between the haplotypes of the six SNPs and breast cancer risk was also performed using a score test developed by Schaid *et al.* (12), as implemented in the computer program HAPLO.SCORE.[7]

For each steroid hormone, subjects were classified into low or high levels based on the pre- or postmenopausal specific median levels. Years of menstruation were calculated for postmenopausal women (age at menopause − age at menarche), and premenopausal women (age at diagnosis − age at menarche for cases, or age at interview − age at menarche for controls).

Results

The characteristics of cases and controls in this study are presented in Table 1. Risk factors identified from this study were consistent with those reported elsewhere (13). Sequence analysis in a subset of 30 subjects revealed eight sequence variants (Table 2). Among these, three are in the promoter region, two are in the coding regions (both are synonymous changes), two are in the 3′UTR, and one is in intron 5.

Five common sequence variants where the frequency of a rare allele was ≥15% [T(-11891)C, C(14206)T, G(25652)A, A(50766)G, G(50995)A], as well as the synonymous change C(33390)G, were genotyped among the 2369 study subjects. All of the SNPs were in Hardy-Weinberg equilibrium in both cases and controls. There is strong LD between all pairs of SNPs, with estimates of D' between 0.89–1.00 (data not shown). The genotype frequencies of these SNPs

[6] Internet address: http://www.wfubmc.edu/genomics.

[7] Internet address: http://www.mayo.edu/statgen for the S-PLUS programming language or http://www.wfubmc.edu/genomics for the R programming language.

Table 1 *Characteristics of case patients and control subjects in Shanghai Breast Cancer Study*

	Case patients (n = 1459)	Control subjects (n = 1556)	*P*
Demographic factors			
Age (years, mean ± SD)	47.79 ± 7.99	47.24 ± 9.11	0.02
Education of high school or higher (%)	45.17	44.34	0.64
Major risk factors			
First degree of relatives with breast cancer (%)	3.7	2.44	0.04
Ever diagnosed with breast fibroadenoma (%)	9.6	5.01	<0.0001
Age at menarche ≤ 14 years (%)	52.84	48.26	0.01
Age at menopause ≥ 52 years (%)[a]	20.96	16.01	0.04
Age at first live birth ≤ 27 years (%)	67.07	71.33	<0.0001
Waist to hip ratio (mean ± SD)	0.81 ± 0.06	0.80 ± 0.06	0.0005
BMI (mean ± SD)[b]	23.53 ± 3.40	23.15 ± 3.39	0.002

[a] Among post-menopausal women.

[b] BMI = weight (in kilograms)/[height (in meters)]^2.

in all of the cases and controls, as well as in pre- or postmenopausal women, are presented in Table 3. There was no statistically significant difference in the allele frequencies between cases and controls for any of the six SNPs in the whole dataset. However, among postmenopausal women, the difference of allele frequencies between cases and controls was found to be statistically significant for the SNP C(14206)T (P = 0.03) and marginally significant for the SNP C(33390)G (P = 0.07). No significant difference of allele frequencies between cases and controls was observed for any of the SNPs in premenopausal women.

One possible explanation for observing this association only in the postmenopausal women may be that these women generally have a longer lifetime exposure to estrogens. If this were true, we would expect to observe a stronger association among women with greater years of menstruation, regardless of menopausal status. To explore this explanation, we identified women at the top quartile of menstruation years in this population (≥34 years). This group included 315 premenopausal and 548 postmenopausal women. Stronger evidence for association was observed for the SNP C(33390)G in this subset of women after adjusting for menopausal status, age, age at first birth, age at menarche, BMI, and family history (Table 4). Women with the risk genotypes of this SNP (CG or GG) had a RR of 2.41 (95% CI, 1.19–4.86; P = 0.01) for breast cancer, compared with women with the nonrisk genotype (CC). Similar results were observed when the analysis was limited to postmenopausal women (data not shown). Among women in the top quartile of menstruation years in this subset (≥36 years), carriers of these risk genotypes of the SNP C(33390)G (CG or GG) had a RR of 3.62 (95% CI, 1.26–10.37; P = 0.01) for breast cancer, compared with women with the nonrisk genotype (CC).

To additionally evaluate the possible modifying effect of endogenous estrogen exposure on the genetic association of *ESR2* variants with breast cancer, we analyzed data on the combination of the risk genotypes (CG or GG) at the SNP C(33390)G and levels of sex steroid hormones separately for post- and premenopausal women. Among the postmenopausal women (Table 5), subjects were at intermediately increased risk for breast cancer if they had either a risk genotype or a high level of each of the steroid sex hormones. In general, the highest risk was observed for those who had both a risk genotype and high levels of sex hormones (testosterone, estradiol, or DHEA-S). Similar results were observed when the analysis was limited to premenopausal women (data not shown in table).

Table 2 *Polymorphisms in the ESR2 gene identified in 30 women from the Shanghai Breast Cancer Study*

SNP[a]	Location[b]	SNP ID	In human-mouse homologous region	Exonic Splicing Enhancer
G(-12214)T	Promoter (-1839)	rs1271572		
G(-11943)A	Promoter (-568)			
T(-11891)C	Promoter (-515)			
C(14206)T	Intron 5 (16bp downstream of e4)			
G(25652)A	exon 6	rs1256049 (V328V)	Yes	SRp55 (degenerate change)
C(33390)G	exon 7	rs1256054 (L392L)	Yes	SC35[c]
A(50766)G	3'UTR			
G(50995)A	3'UTR			SC35 (degenerate change)

[a] The numbers indicate the SNP locations relative to the start codon ATG according to National Center for Biotechnology Information genomic contig NT_026437.

[b] The number indicates the SNP locations relative to the Refseq NM_001437, which contains the first untranslated exon. The positions for the SNPs located in the promoter region are relative to the transcription start site.

[c] Allele C: above background frequency; allele G: below background frequency.

Table 3 *Association of ESR2 polymorphisms and breast cancer risk among pre- and postmenopausal women*

		All women			Premenopausal women			Postmenopausal women		
SNP		Cases n (%)	Controls n (%)	P^a	Cases n (%)	Controls n (%)	P^a	Cases n (%)	Controls n (%)	P^a
T(-11891)C	TT	912 (81.5)	994 (82.1)		617 (82.8)	624 (81.0)		292 (79.1)	368 (84.2)	
	CT	198 (17.7)	206 (17.0)		121 (16.2)	139 (18.1)		75 (20.3)	65 (14.9)	
	CC	9 (0.8)	11 (0.9)	0.78	7 (0.9)	7 (0.9)	0.41	2 (0.5)	4 (0.9)	0.11
C(14206)T	CC	905 (82.3)	991 (83.4)		616 (83.8)	624 (82.5)		286 (79.7)	365 (85.3)	
	CT	185 (16.8)	192 (16.2)		113 (15.4)	128 (16.9)		70 (19.5)	62 (14.5)	
	TT	9 (0.8)	5 (0.4)	0.38	6 (0.8)	4 (0.5)	0.63	3 (0.8)	1 (0.2)	0.03
G(25652)A	GG	480 (43.1)	537 (44.4)		315 (42.5)	340 (44.1)		163 (44.5)	194 (44.7)	
	AG	506 (45.5)	541 (44.8)		339 (45.7)	352 (45.7)		165 (45.1)	188 (43.3)	
	AA	127 (11.4)	131 (10.8)	0.5	88 (11.9)	79 (10.2)	0.34	38 (10.4)	52 (12.0)	0.76
C(33390)G	CC	1032 (93.5)	1145 (94.7)		687 (93.7)	724 (94.0)		340 (92.9)	417 (95.9)	
	CG	69 (6.3)	63 (5.2)		44 (6.0)	46 (6.0)		25 (6.8)	17 (3.9)	
	GG	3 (0.3)	1 (0.1)	0.16	2 (0.3)	0 (0.0)	0.65	1 (0.3)	1 (0.2)	0.07
A(50766)G	AA	346 (31.5)	407 (33.8)		231 (31.6)	256 (33.6)		114 (31.7)	150 (34.4)	
	AG	541 (49.3)	570 (47.4)		359 (49.0)	360 (47.2)		180 (50.0)	207 (47.5)	
	GG	210 (19.1)	226 (18.8)	0.36	142 (19.4)	147 (19.3)	0.56	66 (18.3)	79 (18.1)	0.55
G(50995)A	GG	834 (75.4)	907 (75.6)		553 (74.9)	577 (75.9)		276 (76.0)	327 (75.2)	
	AG	256 (23.1)	267 (22.3)		173 (23.4)	166 (21.8)		83 (22.9)	100 (23.0)	
	AA	16 (1.4)	25 (2.1)	0.84	12 (1.6)	17 (2.2)	0.28	4 (1.1)	8 (1.8)	0.64

[a] P was based on χ^2 test for allele frequency difference between cases and controls.

Table 4 *Association of ESR2 polymorphisms and breast cancer risk among women with greater years of menstruation*[a]

SNP		Number of subjects (%)			
		Cases	Controls	RR[b] (95% CI)	P^b
T(-11891)C	TT	260 (78.1)	245 (81.9)	Reference	
	CT and CC	73 (21.9)	54 (18.1)	1.27 (0.85–1.90)	0.25
C(14206)T	CC	257 (78.6)	246 (83.4)	Reference	
	CT and TT	70 (21.4)	49 (16.6)	1.37 (0.90–2.07)	0.14
G(25652)A	GG	151 (45.5)	129 (43.4)	Reference	
	AG and AA	181 (54.5)	168 (56.6)	0.97 (0.70–1.35)	0.89
C(33390)G	CC	303 (91.3)	284 (95.9)	Reference	
	CG and GG	29 (8.7)	12 (4.1)	2.41 (1.19–4.86)	0.01
A(50766)G	AA	107 (33.2)	96 (32.3)	Reference	
	AG and GG	215 (66.8)	201 (67.7)	1.02 (0.72–1.44)	0.91
G(50995)A	GG	251 (76.1)	225 (75.8)	Reference	
	AG and AA	79 (23.9)	72 (24.2)	1.02 (0.70–1.49)	0.92

[a] Women with years of menstruation above upper quartile (≥34 years).

[b] Adjusted for menopausal status, age, age at first birth, age at menarche, BMI, and family history.

The association between haplotypes of the six SNPs and breast cancer risk was also assessed. Five major haplotypes, with frequencies of at least 1%, were inferred in this population. No association between the haplotypes and breast cancer risk were detected in the whole dataset or among the premenopausal women, using either an omnibus test or individual haplotype. However, a suggestive association between the haplotypes and breast cancer risk was detected using an omnibus test in the postmenopausal women ($P = 0.09$). A significant association ($P = 0.02$) was detected for a specific haplotype [C-T-G-C-G-G of the SNPs T(-11891)C, C(14206)T, G(25652)A, C(33390)G, A(50766)G, and G(50995)A]. This haplotype was more frequent in cases (10%) than in controls (7%).

Discussion

We hypothesize that sequence variants of the ER β gene (*ESR2*) are associated with an increased risk for breast cancer, particularly among women who have a high-level and long-term estrogen exposure. This hypothesis was primarily based on consistent epidemiological evidence and strong biological support for the critical roles of steroid hormones, via activating ERs in the regulation of mammary cell growth and differentiation. Numerous epidemiological studies consistently suggest factors, such as early age at menarche and late age at menopause, that increase the number of menstrual cycles (and therefore lengthen the exposure to estrogens) elevating breast cancer risk (14). Studies from basic biology have shown that estrogens are strong mitogens for mammary cells, and numerous animal studies have demonstrated that estrogens can induce and promote breast cancer, whereas the removal of ovaries or administration of antiestrogens such as tamoxifen can oppose the carcinogenesis process (14, 15). Estrogen action is mediated primarily through binding to ERs (ESRs), which then act as transcriptional factors by binding to specific ER elements of target genes (16, 17), interact with other transcriptional factors such as SP1, AP1, or nuclear factor κB in the absence of DNA binding (18–21), or crosstalk with proteins in other signaling pathways such as raf-mitogen-activated protein kinase signaling and phosphatidylinositol 3′-kinase/Akt signaling pathways via nongenomic effects (4, 22).

Most of the previous studies have focused on ER α (*ESR1*), and

Table 5 *Joint effect of estrogens and ESR2 sequence variants on breast cancer risk among postmenopausal women*

Hormones[a]	C(33390)G	Cases	Controls	RR[b] (95% CI)	P^b
Testosterone					
Low	CC	80	222	Reference	
High	CC	97	181	2.11 (1.37–3.23)	0.0007
Low	CG and GG	6	9	2.38 (0.73–7.78)	0.15
High	CG and GG	9	6	6.63 (1.92–22.89)	0.003
Estradiol					
Low	CC	100	200	Reference	
High	CC	77	203	0.96 (0.64–1.45)	0.84
Low	CG and GG	3	7	0.97 (0.22–4.39)	0.97
High	CG and GG	12	8	3.96 (1.36–11.56)	0.01
Estrone					
Low	CC	77	235	Reference	
High	CC	99	164	1.59 (1.07–2.37)	0.02
Low	CG and GG	8	8	3.66 (1.21–11.05)	0.02
High	CG and GG	7	7	2.18 (0.62–7.59)	0.22
Estrone-S					
Low	CC	72	218	Reference	
High	CC	104	185	1.63 (1.10–2.42)	0.01
Low	CG and GG	7	7	3.62 (1.08–12.05)	0.04
High	CG and GG	8	8	3.12 (0.93–10.52)	0.06
DHEA-S					
Low	CC	74	213	Reference	
High	CC	101	184	1.71 (1.15–2.55)	0.008
Low	CG and GG	4	8	1.78 (0.47–6.74)	0.40
High	CG and GG	11	7	5.18 (1.68–15.94)	0.004
SHBG					
High	CC	75	219	Reference	
Low	CC	104	184	1.24 (0.82–1.86)	0.30
High	CG and GG	5	6	3.59 (0.95–13.65)	0.06
Low	CG and GG	10	9	2.55 (0.83–7.86)	0.10

[a] Using post-menopausal specific median to divide into high or low steroid hormone levels.

[b] Adjusted for age, age at first birth, age at menarche, age at menopause, BMI, and batch for hormone measurement.

several epidemiological studies have shown that genetic polymorphisms in the *ESR1* gene may be associated with breast cancer risk (9, 10, 23). The importance of *ESR2* in breast cancer has not been recognized until very recently (24). *ESR2* and *ESR1* have distinct cellular distributions, regulate separate sets of genes, and oppose each other's actions when regulating some genes. *ESR2* is widely expressed in both normal and malignant breast, and there are proliferating cells in the breast that express *ESR2*. Considering the fact that only a subset of women who have a high level or long-term estrogen exposure develop breast cancer, it is possible that the risk of breast cancer may be modified by polymorphisms in genes, such as *ESR2*, that are involved in the regulation of estrogen effect in breast tissues. The results of this study appear to be consistent with this hypothesis.

However, caution should be taken when interpreting our findings. In particular, type I errors are a concern in association studies of genetic polymorphisms, given the fact that multiple comparisons are often made. It is difficult, however, to adjust the *P*s for multiple comparisons in this study, because these sequence variants are not independent due to strong LD. The hormone levels are not independent either, with significant interclass correlation coefficients. In addition, because the functional relevance of the polymorphisms evaluated is unknown, additional epidemiological studies and functional studies are needed.

Our results are also susceptible to multiple sources of measurement variation associated with steroid hormones. However, we have paid particular attention to minimize such random variation by using an individual matched study design that enhances the comparability between cases and controls, as well as by minimizing the variability of laboratory testing from batch to batch (between-assay variation).

At present, the mechanism by which the SNP C(33390)G interacts with hormones and affects breast cancer risk is unknown. The observed association may be due to a causal effect of the SNP or through other as yet unknown flanking sequence variants in LD with this SNP. Although C(33390)G is a silent synonymous change, results from multiple studies have shown that synonymous changes may inactivate genes by inducing the splicing machinery to skip the exons (25–28). In fact, the SNP C(33390)G is located in an exonic splicing enhancer motif (in the binding site for splice factor SC35), and, thus, may affect the accuracy and efficiency of *ESR2* pre-mRNA splicing. The C allele of C(33390)G was preferentially found in the SC35 enhancer element at this particular position (has above background frequency in SC35 consensus sequence), whereas the G allele had below background frequency and may have an adverse effect on the splicing efficiency of exon 7. In addition, we observed an association between G allele carriers of the SNP and self-reported fibroadenoma risk in our study (10.5% in 172 cases *versus* 5.5% in 2140 controls; RR, 2.0; 95% CI, 1.19–3.38; data not shown), providing additional support for a potential functional impact of this variant.

It is worth noting that the joint effect of higher estrogen levels and *ESR2* sequence variants was consistently observed for multiple estrogens and estrogen-related steroids, including DHEA-S. One of the major sources of steroid hormones in women is the synthesis in intracrine tissues from the inactive precursors DHEA and DHEA-S of adrenal origin; this intracrine synthesis is especially important after menopause. The importance of steroid hormones from an intracrine source is illustrated by several clinical trials that describe the benefits of aromatase (which converts androgens to estrones) inhibitors in the treatment of hormone-responsive metastasis (ER+) as well as primary breast cancers in women (29–31). Higher circulating DHEA as well as its sulfate conjugate (DHEA-S) provides the peripheral target tissues more estrogen precursor and may lead to higher local estrogen levels. Higher estrogen levels, when conjugated with ESR that has increased activity, can additionally increase the risk of breast cancer. Again, caution should be exercised when interpreting the interaction effect, especially given the small sample size for this particular part of the study.

Our results, if confirmed, may have significant implications in the prevention of breast cancer. Women with the "high-risk" ESR2 sequence variants could reduce their risk for breast cancer by reducing their estrogen levels.

Acknowledgments

We thank all of the study subjects who participated in this study.

References

1. Jin, F., Devesa, S. S., Chow, W. H., Zheng, W., Ji, B. T., Fraumeni, J. F., Jr., and Gao, Y. T. Cancer incidence trends in urban shanghai, 1972–1994: an update. Int. J. Cancer, *83:* 435–440, 1999.
2. Gao, Y. T., Shu, X. O., Dai, Q., Potter, J. D., Brinton, L. A., Wen, W., Sellers, T. A., Kushi, L. H., Ruan, Z., Bostick, R. M., Jin, F., and Zheng, W. Association of menstrual and reproductive factors with breast cancer risk: results from the Shanghai Breast Cancer Study. Int. J. Cancer, *87:* 295–300, 2000.
3. Yu, H., Shu, X. O., Shi, R., Dai, Q., Jin, F., Gao, Y. T., Li, B. D., and Zheng, W. Plasma sex steroid hormones and breast cancer risk in Chinese women. Int. J. Cancer, *105:* 92–97, 2003.
4. Sanchez, R., Nguyen, D., Rocha, W., White, J. H., and Mader, S. Diversity in the mechanisms of gene regulation by estrogen receptors. Bioessays, *24:* 244–254, 2002.
5. Speirs, V., Adams, I. P., Walton, D. S., and Atkin, S. L. Identification of wild-type and exon 5 deletion variants of estrogen receptor β in normal human mammary gland. J. Clin. Endocr. Metab., *85:* 1601–1605, 2000.
6. Poola, I., Abraham, J., and Liu, A. Estrogen receptor β splice variant mRNAs are differentially altered during breast carcinogenesis. J. Steroid. Biochem. Mol. Biol., *82:* 169–179, 2002.
7. de Jong, M. M., Nolte, I. M. te Meerman, G. J., van der Graaf, W. T., Oosterwijk, J. C., Kleibeuker, J. H., Schaapveld, M., and de Vries, E. G. Genes other than BRCA1 and BRCA2 involved in breast cancer susceptibility. J. Med. Genet., *39:* 225–242, 2002.
8. Zheng, W., Xie, D. W., Jin, F., Cheng, J. R., Dai, Q., Wen, W. Q., Shu, X. O., and Gao, Y. T. Genetic polymorphism of cytochrome P450-1B1 and risk of breast cancer. Cancer Epidemiol. Biomark. Prev., *9:* 147–150, 2000.
9. Cai, Q., Shu, X-O., Dai, Q., Wen, W., Cheng, J-R, Gao, Y-T., and Zheng, W. Genetic polymorphisms of the estrogen receptor-α gene and risk of breast cancer: Results from the Shanghai Breast Cancer Study. Cancer Epidemiol. Biomark. Prev., in press, 2003.
10. Cai, Q. Y., Gao, Y. T., Smith, J., Wen, W., Shu, X. O., Jin, F., and Zheng, W. A GT dinucleotide repeat polymorphism upstream of the estrogen receptor-α gene in relation to breast cancer risk. Cancer Res., in press, 2003.
11. Weir, B. S. Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sunderland, MA: Sinauer Association, Inc. Publishers, 1996.
12. Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., and Poland, G. A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am. J.Hum. Genet., *70:* 425–434, 2002.
13. Nathanson, K. L., and Weber, B. L. Breast cancer. *In:* R. A. King, J. I. Rotter, and A. G. Motulsky (eds.), The Genetic Basis of Common Diseases, pp. 670–699. New York: Oxford University Press, Inc., 2002.
14. Hulka, B. S., and Moorman, P. G. Breast cancer: hormones and other risk factors. Maturitas, *38:* 103–113, 2001.
15. Vignon, F., Bouton, M. M., and Rochefort, H. Antiestrogens inhibit the mitogenic effect of growth factors on breast cancer cells in the total absence of estrogens. Biochem. Biophys. Res. Commun., *146:* 1502–1508, 1987.
16. Tsai, M. J., and O'Malley, B. W. Molecular mechanisms of action of steroid/thyroid receptor superfamily members. Annu. Rev. Biochem., *63:* 451–486, 1994.
17. Dickson, R. B., and Stancel, G. M. Estrogen receptor-mediated processes in normal and cancer cells. J. Natl. Cancer Inst. Monogr., 135–145, 2000.
18. Batistuzzo, d. M., Sr., Krey, G., Hihi, A. K., and Wahli, W. Functional interactions between the estrogen receptor and the transcription activator Sp1 regulate the estrogen-dependent transcriptional activity of the vitellogenin A1 io promoter. J. Biol. Chem., *272:* 18250–18260, 1997.
19. Galien, R., Evans, H. F., and Garcia, T. Involvement of CCAAT/enhancer-binding protein and nuclear factor-κ B binding sites in interleukin-6 promoter inhibition by estrogens. Mol. Endocrinol., *10:* 713–722, 1996.
20. Paech, K., Webb, P., Kuiper, G. G., Nilsson, S., Gustafsson, J., Kushner, P. J., and Scanlan, T. S. Differential ligand activation of estrogen receptors ERα and ERβ at AP1 sites. Science (Wash. DC), *277:* 1508–1510, 1997.
21. Porter, W., Saville, B., Hoivik, D., and Safe, S. Functional synergy between the transcription factor Sp1 and the estrogen receptor. Mol. Endocrinol., *11:* 1569–1580, 1997.
22. Lee, A. V., Cui, X., and Oesterreich, S. Cross-talk among estrogen receptor, epidermal growth factor, and insulin-like growth factor signaling in breast cancer. Clin. Cancer Res., *7:* 4429s–4435s, 2001.

23. Kang, H. J., Kim, S. W., Kim, H. J., Ahn, S. J., Bae, J. Y., Park, S. K., Kang, D., Hirvonen, A., Choe, K. J., and Noh, D. Y. Polymorphisms in the estrogen receptor-α gene and breast cancer risk. Cancer Lett., *178:* 175–180, 2002.

24. Palmieri, C., Cheng, G. J., Saji, S., Zelada-Hedman, M., Warri, A., Weihua, Z., Van Noorden, S., Wahlstrom, T., Coombes, R. C., Warner, M., and Gustafsson, J. A. Estrogen receptor β in breast cancer. Endocr. Relat. Cancer, *9:* 1–13, 2002.

25. Cartegni, L., Chew, S. L., and Krainer, A. R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat. Rev. Genet., *3:* 285–298, 2002.

26. Caputi, M., Kendzior, R. J., Jr., and Beemon, K. L. A nonsense mutation in the fibrillin-1 gene of a Marfan syndrome patient induces NMD and disrupts an exonic splicing enhancer. Genes Dev., *16:* 1754–1759, 2002.

27. Cartegni, L., and Krainer, A. R. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. Nat. Genet., *30:* 377–384, 2002.

28. Fackenthal, J. D., Cartegni, L., Krainer, A. R., and Olopade, O. I. BRCA2 T2722R is a deleterious allele that causes exon skipping. Am. J. Hum. Genet., *71:* 625–631, 2002.

29. Lake, D. E., and Hudis, C. Aromatase inhibitors in breast cancer: an update. Cancer Control, *9:* 490–498, 2002.

30. Coleman, R. E. Current and future status of adjuvant therapy for breast cancer. Cancer (Phila.), *97:* 880–886, 2003.

31. Cuzick, J., Powles, T., Veronesi, U., Forbes, J., Edwards, R., Ashley, S., and Boyle, P. Overview of the main outcomes in breast-cancer prevention trials. Lancet, *361:* 296–300, 2003.

Appendix C

Genomic structure and the expression patterns of *RNASEL*

Baoli Chang[1], Amy Tolin[1], Fernando Segade[1], Robert Silverman[2], William B. Isaacs[3], Jianfeng Xu[1]

[1]Center for Human Genomics, Wake Forest University School of Medicine, Winston-Salem, NC; [2]Department of Cancer Biology, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, OH; [3]Brady Urological Institute, Johns Hopkins Medical Institutions, Baltimore, MD

RNase L is a single-strand-specific endonuclease that functions as the terminal enzyme component of the interferon-inducible 2-5A synthetase/RNAase L antiviral pathway. Recently, the gene encoding RNase L, *RNASEL*, was identified as a strong candidate for the prostate cancer susceptibility gene at the HPC1 locus. However, neither the public databases nor published studies have fully described and annotated the complete genomic structure of the *RNASEL* gene. In this study, we performed 1) Northern-blot analysis to estimate the approximate size of the *RNASEL* transcript, as well as the expression levels in various tissues, 2) RT-PCR to quantify the number of exons, 3) 5'- and 3'-RACE to define both ends of the *RNASEL* transcript. With a "*RNASEL*-unique" sequence for our northern probe, we were able to avoid cross-hybridization to homologous proteins and identify single transcript specie at ~4.4kb. In addition, the highest expression was found to be in the prostate, followed by testis and small intestine. The rest of the tissues have very low RNASEL expression levels. To determine the total number of exons, as well as the 5'- and 3'- end of the RNASEL transcript, we carried out RT-PCR, 5'- and 3'-RACE. A 7-exon model, including an untranslated 1st exon which is not annotated by NCBI, was suggested both by our experimental data and computational annotation based on an EST database search. Interestingly, comparative genomic analyses between human *RNASEL* and mouse *Rnasel* loci identified several homologous regions in the proximal promoter and exon 1, which indicate a common transcriptional regulation mechanisms between the two species.

[CANCER RESEARCH 64, ●●●–●●●, April 15, 2004]

Sequence Variants of *Toll-Like Receptor 4* Are Associated with Prostate Cancer Risk: Results from the CAncer Prostate in Sweden Study

S. Lilly Zheng,[1] Katarina Augustsson-Bälter,[2] Baoli Chang,[1] Maria Hedelin,[2] Liwu Li,[1] Hans-Olov Adami,[2] Jeanette Bensen,[1] Ge Li,[1] Jan-Erik Johnasson,[4] Aubrey R. Turner,[1] Tamara S. Adams,[1] Deborah A. Meyers,[1] William B. Isaacs,[5] Jianfeng Xu,[1] and Henrik Grönberg[3]

[1]Center for Human Genomics, Wake Forest University School of Medicine, Winston-Salem, North Carolina; [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [3]Department of Radiation Sciences, Oncology, University of Umeå, Umeå, Sweden; [4]Department of Urology and Clinical Medicine, Örebro University Hospital, Sweden, and Regional Oncological Center, University Hospital, Uppsala, Sweden; and [5]Department of Urology, Johns Hopkins Medical Institutions, Baltimore, Maryland

ABSTRACT

Inflammation has been implicated as an etiological factor in several human cancers. Growing evidence suggests that chronic inflammation may also play a role in the etiology of prostate cancer. Considering that genetic susceptibility is a major risk factor for this disease, we hypothesize that sequence variants in genes that regulate inflammation may modify individual susceptibility to prostate cancer. The lipopolysaccharide receptor Toll-like receptor 4 (*TLR4*) is a central player in the signaling pathways of the innate immune response to infection by Gram-negative bacteria and is an important candidate inflammatory gene. We performed a systematic genetic analysis of *TLR4* sequence variants by evaluating eight single-nucleotide polymorphisms that span the entire gene among 1383 newly diagnosed prostate cancer patients and 780 age- and residence-matched controls in Sweden. We found an association between a sequence variant (11381G/C) in the 3′-untranslated region of the *TLR4* gene and prostate cancer risk. The frequency of the variant genotypes (CG or CC) was significantly higher in the patients (24.1%) than in the controls (19.7%; $P = 0.02$). The frequency of risk genotypes among patients diagnosed before the age of 65 years was even higher (26.3%). Compared with men who had the wild-type genotype of this single-nucleotide polymorphism (GG), those with GC or CC genotypes had a 26% increased risk for prostate cancer (odds ratio, 1.26; 95% confidence interval, 1.01–1.57) and 39% increased risk increased risk for early onset prostate cancer (before age 65 years; odds ratio, 1.39; 95% confidence interval, 1.02–1.91). The risk attributable to this variant for prostate cancer in Sweden was estimated to be 4.9%. Although the biological mechanism of the observed association remains to be elucidated, our finding supports a role for a bacteria-associated response pathway, possibly acting via inflammation, in the development of prostate cancer.

AQ: A

INTRODUCTION

Chronic or recurrent inflammation is known to play a causative role in the development of many human cancers, including cancers of the liver, esophagus, stomach, large intestine, and urinary bladder (1). Inflammatory changes have long been recognized in prostate tissues, leading to speculation that chronic inflammation might contribute to prostate cancer development (2). The role of prostate inflammation in prostate cancer was strongly implicated in the recently proposed theory that proliferative inflammatory atrophy serves as a precursor to prostatic intraepithelial neoplasia and to prostate cancer (3, 4). The identification of two candidate prostate cancer susceptibility genes (*RNASEL* and *MSR1*) that encode proteins with critical functions in host responses to infections provides additional support for a role of inflammation in prostate cancer development (5, 6).

Chronic infection and inflammatory processes that may lead to tumorigenesis are mediated in part through recognition of various stimuli by Toll-like receptors (TLRs; Ref. 7). Among these TLRs, *TLR4* recognizes Gram-negative bacterial products, including lipopolysaccharide (8), the antitumor compound taxol in the mouse but not in humans (9), and human heat shock protein 60 (10). Differential activation of *TLR4* by this array of naturally occurring or synthetic ligands subsequently induces distinct downstream processes, including the expression of inflammatory genes as well as regulation of cell growth/apoptosis. Conceivably, improper regulation or compromised function of *TLR4* may contribute to various inflammatory diseases, including cancer.

On the basis of the potential importance of inflammation and inflammatory genes in prostate cancer development, we hypothesized that sequence variants of *TLR4* are associated with prostate cancer susceptibility. To test this hypothesis, we performed a systematic genetic analysis in a large population-based prostate cancer case–control study in Sweden.

MATERIALS AND METHODS

Study Population. The cases studied came from the large-scale, population-based case–control study CAncer Prostate in Sweden (CAPS). The case participants were recruited from four of the six regional cancer registries that cover the entire population of Sweden. Each of these four registries serves one health care region (Northern, Central, Stockholm, and South Eastern); the four registries altogether encompass ~6 millions inhabitants (67% of Sweden's population). Reporting of newly diagnosed cancer cases to the registries is required by law for both the attending physician and pathologist; therefore, the registries include almost 100% of all cancers diagnosed in Sweden.

In the CAPS study, the source-person-time comprises men living in the area of Örebro and the northern part of Sweden (Västernorrland, Jämtland, Västerbotten, and Norrbotten) from January 1, 2001, as well as men living in the areas of Västmanland, Södermanland, Gävleborg, Dalarna, Värmland, and Uppland from July 1, 2001, until September 2002 (except for Jämtland and the county of Lycksele in Västerbotten, where the source-person-time ended March 1, 2002). The source-person-time was divided into two age-specific study bases. The first study base included men age 35–65 years of age living in all of the regions mentioned above. The second study base included men 66–79 years of age living only in the areas of Örebro, Västmanland, Södermanland, and the northern part of Sweden.

The inclusion criterion for cases in CAPS was pathologically or cytologically verified adenocarcinoma of the prostate (ICDO = C61). After receiving AQ: B notification of a new case, the administrator at the regional cancer registry mailed a letter to the treating physician informing him or her about the study. The physician was asked to indicate whether the patient was able to participate in the study. If so, the physician mailed a letter to the patient to introduce the study and asked him to send a reply letter to the administrator at the cancer registry. After approval from both the physician and the patient, the study secretaries sent a questionnaire and a kit with tubes for blood sampling to the eligible case. The self-administered questionnaire included questions concern-

Received 10/18/03; revised 2/6/04; accepted 2/9/04.

Grant support: Grants from the Swedish Cancer Foundation and Spear Grant from the Umeå University Hospital (Umeå, Sweden). Also partially funded by the Center for Human Genomics at Wake Forest University School of Medicine.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Jianfeng Xu, Wake Forest University School of Medicine, Medical Center Blvd, Winston-Salem, NC 27157. Phone: (336) 713-7500; Fax: (336) 713-7566; E-mail: jxu@wfubmc.edu.

DOI: 10.1158/0008-5472.CAN-03-3280

ing such items as diet (validated food questionnaire), family history, smoking, and physical activity.

In total, 1961 prostate cancer cases were invited to participate; of these, 1444 (73.6%) agreed to participate by donating a blood sample and/or answering the questionnaire. DNA was available for 1383 (95.8%) of the cases that participated. Clinical data that were not included in the Cancer Registry were Fn6 obtained from the National Prostate Cancer Registry.[6] The cases were linked AQ: C to the National Prostate Cancer Registry, and clinical information such as tumor-node-metastasis (TNM) stage, Gleason sum, prostate-specific antigen (PSA) level at the time of diagnosis, means of diagnosis, and primary treatment were obtained for 95.3% of the cases. The cases were thereafter classified as either localized (T_{1-2} and N_0/N_X and M_0/M_X; grade I-II; Gleason sum, 2–7; AQ: D and PSA <100) or advanced (prone to progressive disease; $T_{3/4}$ or N+ or M+, or grade III, or Gleason sum of 8–10, or PSA >100).

Control subjects were randomly selected from the updated Swedish Population Registry based on frequency matching to the expected age distribution (within 5 years) and geographic origin of the cases. After the controls were identified, a letter of introduction to the study was mailed to each control. Three to 4 weeks later, the same questionnaire and blood sampling kit that were used for the cases were mailed to the controls. Of the 1697 randomly selected controls invited, 866 (52.0%) agreed to participate by donating a blood sample and/or completing a questionnaire. DNA was available for 780 (90.9%) of the controls that participated. Eight potential control subjects were excluded after linkage to the National Cancer Registry revealed that they had a diagnosis of prostate cancer before inclusion.

To improve the response rate, cases and controls were recontacted three times; after 1–2 weeks with a follow-up letter, after 6–8 weeks with a new questionnaire and blood draw kit, and after ~12 weeks with a phone call. The T1 clinical characteristics of the study subjects are presented in Table 1. Mean age (age at diagnosis for cases and age at inclusion for controls) for the cases and controls were 66.60 and 67.90 years, respectively.

Table 1 *Characteristics of study subjects in the CAncer Prostate in Sweden 1 study*

	Age at diagnosis			
	45–65 years		66–80 years	
	n	%	n	%
T stage				
T_0	3	0.45	3	0.46
T_1	259	39.24	200	30.96
T_2	234	35.45	208	32.20
T_3	144	21.82	207	32.04
T_4	20	3.03	28	4.33
Missing	44		33	
N stage				
N_0	157	85.79	66	77.65
N_1	26	14.21	19	22.35
Missing	521		594	
M stage				
M_0	328	83.04	250	79.62
M_1	67	16.96	64	20.38
Missing	309		365	
Gleason score				
≤4	29	4.67	24	4.21
5	79	12.72	63	11.05
6	236	38.00	185	32.46
7	177	28.50	188	32.98
8	53	8.53	70	12.28
9	40	6.44	37	6.49
10	7	1.13	3	0.53
Missing	83		109	
Differential				
GI	35	13.51	34	13.23
GII	154	59.46	151	58.75
GIII	70	27.03	72	28.01
Missing	445		422	
PSA[a] levels				
<4	49	7.46	26	4.07
4–9.99	264	40.18	169	26.45
10–19.99	139	21.16	152	23.79
20–49.99	92	14.00	128	20.03
50–99.99	49	7.46	77	12.05
≥100	64	9.74	87	13.62
Missing	47		40	

[a] PSA, prostate-specific antigen.

This study was approved by the Ethical Committees at the Karolinska Institutet and at Umeå University. Written informed consent was obtained from each subject.

Genotyping Methods. All participants in this study were instructed to donate blood (4 × 10 ml) at the nearest health clinic or hospital. Samples were thereafter kept at room temperature and sent by overnight mail to the Medical Biobank at Umeå University. After arrival at the Biobank, leukocytes, erythrocytes, plasma, and serum were separated into different tubes. Samples were stored at −70°C until time for analysis. DNA samples were extracted from whole blood by standard methods and were shipped from Umeå, Sweden to the genotyping laboratory in the Center for Human Genomics at Wake Forest AQ: E University. Each DNA plate contained 2 CEPH controls, a water blank, and blinded internal replicates. Researchers at Wake Forest University were blinded to case status. Genotyping was performed with the MassARRAY system (SEQUENOM). SpectraDesign software was used to generate the primers. Primer sequences and PCR conditions for these sequence variants are Fn7 available at the authors' website.[7]

Selection of *TLR4* single-nucleotide polymorphisms (SNPs). The genomic structure of the *TLR4* gene has not been completely elucidated. The gene is ~11.5 kb and is composed of four exons. There are four known *TLR4* transcript isoforms (A, B, C, and D) that result from alternative splicing of the four exons and different translation start sites. The poly(A) tail has been annotated only for isoform D. Our goal in this study was to evaluate common haplotypes of *TLR4* sequence variants with use of a limited number of SNPs. To achieve this goal, we first selected a subset of reported SNPs from the National Center for Biotechnology Information (NCBI) dbSNP and IIPGA AQ: F databases, using the criteria of minor allele frequencies ≥5% and a density of 1 SNP/kb across the *TLR4* targeted genomic region, including 2 kb of the promoter, all exons, introns, and the predicted 3′-untranslated region (UTR). We also selected a subset of functional and coding SNPs regardless of the F1 frequency and density. A total of 18 SNPs were selected (Fig. 1), including 4 reported nonsynonymous changes (D299G, V310G, E474K, and Q510H, defined based on isoform A). The previously described nonsynonymous change T399I was not selected because it has been reported to be in strong linkage disequilibrium (LD) with the nonsynonymous change D299G (11). We then genotyped these SNPs in a subset of 96 control subjects. Nine of these SNPs, including E474K and Q510H, were monomorphic in these 96 subjects. The nonsynonymous change V310G was observed only once. The remaining eight SNPs were observed multiple times and thus were selected for genotyping among all case and control subjects whose DNA samples were available at the time of this study.

***TLR4* RNA Analyses.** The 3′-UTR boundary of *TLR4* mRNA was determined by 3′ rapid amplification of cDNA end (RACE) on a prostate poly(A)+ RNA template. We used the SMART RACE cDNA amplification kit (Clon- AQ: G tech) according to the manufacturer's recommendations. The gene-specific primer used in the RACE reaction was 5′-ggatccctcccctgtacccttctcact-gccaggag-3′.

Statistical Analysis. Hardy–Weinberg equilibrium tests for each of the eight sequence variants and pair-wise LD tests for all pairs of the eight sequence variants were performed with the Fisher probability test statistic, as described by Weir (12). For each test, 10,000 permutations were performed, and the test statistic of each replicate was calculated. Empirical *P*s for each test were estimated as the proportion of replicates found to be as probable as or less probable than the observed data, as implemented in the software package Genetic Data Analysis. The estimates of pair-wise LD (*D*′) were calculated with the SAS/Genetics computer software package.

Allele frequency differences between the two groups were tested for each SNP by the χ^2 test with 1 degree of freedom (*df*). Genotype frequency differences were also tested by the χ^2 test with 2 degrees of freedom. Both tests were performed with the SAS/Genetics computer program. Odds ratios (ORs) of prostate cancer for the variant-allele carriers (homozygous and heterozygous) *versus* homozygous wild-type allele carriers were estimated by unconditional logistic regression and adjusted for age and geographic regions. Attributable risk was estimated using the formula: 100% × $p(\text{OR} - 1)/[p(\text{OR} - 1) + 1]$, where p is the prevalence of risk genotypes in a population (13).

[6] http://www.roc.SE.

[7] www.wfubmc.edu/genomics.

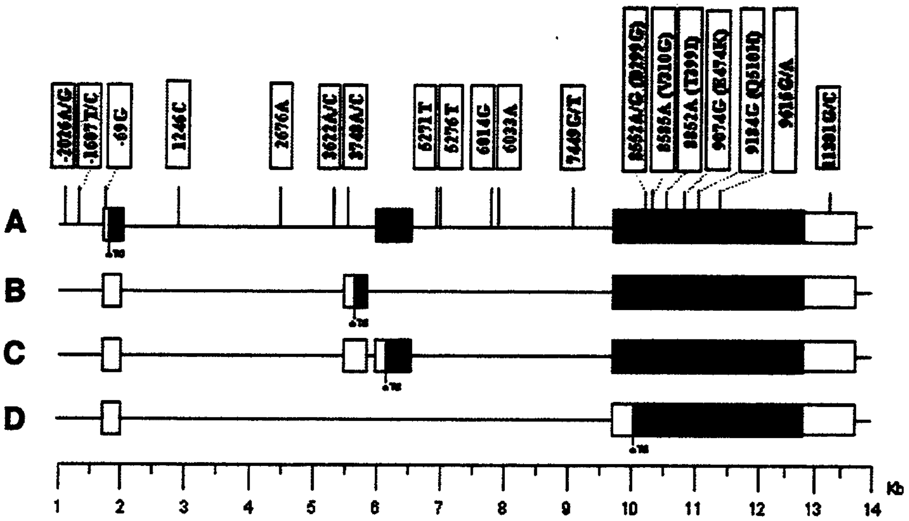

Fig. 1. Toll-like receptor 4 (*TLR4*) genomic structure, with four known transcript isoforms (*A*, *B*, *C*, and *D*). Sequence variants are indicated at the *top* of the figure. All variants shown were screened in 96 controls. *Shaded variants* were genotyped in all study subjects.

Haplotype frequency was estimated with the new statistical method proposed by Stephens *et al.* (14), as implemented in the computer program Fn8 PHASE.[8] When several runs were performed with different values for the seed of the random number generator, the goodness-of-fit values were similar among the different runs. Association between the haplotypes and prostate cancer risk was performed with a score test developed by Schaid *et al.* (15), as Fn9 implemented in the computer program HAPLO.SCORE.[9]

RESULTS

Eight SNPs, including one nonsynonymous change (D299G), were genotyped in all study subjects, including 1383 cases and 780 controls. Because the SNP 7449G/T (rs2149356) significantly deviated from Hardy–Weinberg equilibrium in both cases ($P < 0.0001$) and controls ($P = 0.0006$), it was removed from further statistical analysis. The remaining seven SNPs were in Hardy–Weinberg equilibrium among cases and controls, respectively (all $P > 0.05$). These SNPs were in strong LD: most of the pair-wise D' estimates were 1.0 with the lowest one being 0.63.

Testing for allele frequency differences between cases and controls revealed one SNP (11381G/C) with a marginally significant differ- T2 ence between cases and controls (Table 2). The frequency of the C allele for the SNP 11381G/C was 0.13 and 0.11 in cases and controls, respectively ($P = 0.05$). The genotype frequencies of CC or CG for this SNP were 24.1% and 19.7% in cases and controls, respectively. The difference in genotype frequencies between cases and controls was larger ($\chi^2 = 7.77$; *df*, 2; $P = 0.02$). In logistic regression analyses T3 (Table 3), men who had the genotypes CC or CG had a 26% increased risk (OR = 1.26, 95% confidence interval, 1.01–1.57) for prostate cancer compared with men with the genotype GG for this SNP (adjusted for age and geographic regions). Furthermore, men who had the genotypes CC or CG were at a 39% increased risk (OR = 1.39; 95% confidence interval, 1.02–1.91) for an early diagnosis of prostate cancer. The high-risk genotypes of SNP 11381G/C were present in 19.7% of control individuals, 24.1% of all cases, and 26.3% of cases diagnosed at age <65 years. The overall proportion of prostate cancer risk in this population that was attributable to the risk genotypes was 4.9%.

There was no statistically significant association between the other six SNPs and prostate cancer risk. The frequency of the variant allele G of the nonsynonymous change D299G was similar in cases (0.051) and in controls (0.057; $P = 0.31$). Among the prostate cancer patients only, no association was found between any of the SNPs and clinical characteristics such as age of diagnosis (<65 *versus* ≥65 years), Gleason score (<8 and ≥8), or PSA levels (data not shown). The difference in the genotype frequency of 11381G/C between cases with a Gleason score <8 (24.2%) and ≥8 (23.8%) was not statistically significant. The age-adjusted means for total PSA levels were 16.0, 99.1, and 83.7 among the cases with the CC, CG, and GG genotypes of this SNP, respectively ($P = 0.18$).

Haplotype analysis did not provide additional support for an association between *TLR4* haplotypes and prostate cancer risk (data not shown). There were seven inferred haplotypes in this population based on these seven SNPs. The overall frequencies of these haplotypes were not significantly different ($\chi^2 = 6.52$; *df*, 6; $P = 0.37$). The haplotype containing the C allele of SNP S61381 (haplotype A-T-A-A-A-G-C of SNPs −2026A/G, −1607T/C, 3622A/C, 3748A/C, D299G, 9615G/A, and 11381G/C) was present at a higher frequency in cases (12.43%) than in controls (10.35%; $P = 0.04$).

Because there has been some controversy regarding the boundary of the 3′-UTR for the *TLR4* gene, we performed a 3′ RACE experiment to further define the 3′ end of *TLR4* mRNA expressed in prostate. Our 3′ RACE results indicated that the 3′-UTR of *TLR4* is 1691 bp longer than the 1127-bp NCBI-annotated 3′-UTR, with a total length of 2818 bp. These result indicate that the last SNP in our association study, 11381G/C, is actually located in the middle of the *TLR4* 3′-UTR.

DISCUSSION

There is growing evidence that chronic inflammation may play a role in the development of cancer within several organs, including the prostate. We are in the process of systematically evaluating possible associations between sequence variants of numerous inflammatory genes and prostate cancer risk. *TLR4* is a central player in the signaling pathways that control the innate immune response and was selected to be among the first group of genes to be evaluated. We hypothesized that variations in *TLR4* expression and function may be associated with individual susceptibility to cancer. To test this hypothesis, we performed a systematic genetic analysis of *TLR4* sequence variants in ~2100 men with or without prostate cancer in Sweden. We found a significant association between a sequence variant in the 3′-UTR of the *TLR4* gene. This variant conferred a 26%

[8] http://www.stats.ox.ac.uk/mathgen/software.html.

[9] http://www.mayo.edu/statgen for the S-PLUS programming language; http://www.wfubmc.edu/genomics for the R programming language.

Table 2 *Genotype frequency of Toll-like receptor 4 (TLR4) single-2 nucleotide polymorphisms*

				Subjects, *n* (%)		
SNP[a] (dbSNP)	Location	Amino acid	Genotype	Cases	Controls	*P*[c]
−2026A/G	Promoter		AA	625 (45.45)	341 (43.94)	
			AG	596 (43.35)	354 (45.62)	
			GG	154 (11.20)	81 (10.44)	0.8
−1607T/C	Promoter		TT	991 (72.07)	571 (73.39)	
			TC	350 (25.45)	194 (24.94)	
			CC	34 (2.47)	13 (1.67)	0.34
3622A/C	Intron		AA	1228 (90.56)	680 (88.66)	
			AC	128 (9.44)	82 (10.69)	
			CC	0 (0.00)	5 (0.65)	0.07
3748A/C	Intron		AA	1327 (96.02)	747 (96.02)	
			AC	53 (3.84)	31 (3.98)	
			CC	2 (0.14)	0 (0.00)	0.87
7449G/T (rs2149356)	Intron		GG	603 (51.89)	331 (52.62)	
			TG	423 (36.40)	224 (35.61)	
			TT	136 (11.70)	74 (11.76)	0.83
8552A/G (rs4986790)	Coding	D299G	AA	1241 (90.06)	693 (89.19)	
			AG	136 (9.87)	79 (10.17)	
			GG	1 (0.07)	5 (0.64)	0.31
9615G/A (rs5030721)	Coding		GG	1355 (98.12)	766 (98.21)	
			AG	25 (1.81)	14 (1.79)	
			AA	1 (0.07)	0 (0.00)	0.79
11381G/C	3′-UTR		GG	1047 (75.87)	625 (80.33)	
			CG	318 (23.04)	141 (18.12)	
			CC	15 (1.09)	12 (1.54)	0.05

[a] SNP, single-nucleotide polymorphism; UTR, untranslated region.
[b] From ATG of Isoform A (NM_138554.1)
[c] Test for allele frequencies.

increased risk of prostate cancer in the study population overall and a 39% increased risk of being diagnosed with prostate cancer before age 65. The risk of prostate cancer in Sweden attributable to this variant was estimated to be 4.9%. Our study represents the first comprehensive evaluation of association between sequence variants of the *TLR4* gene and cancer susceptibility.

Although the observed ~1.3–1.4-fold increase in risk is modest, this is probably consistent with the magnitude of risk that we expect to observe for such a heterogeneous disease. Because genes in multiple pathways (such as androgen metabolism, growth factor, phase I and II detoxification, DNA repair, and inflammation) may alter the risk for prostate cancer, each individual gene is likely to contribute only a modest risk. This phenomenon is observed in other complex diseases, as reported in the most recent meta-analysis of genetic association studies in complex diseases (16). After evaluating 301 published studies that attempted to replicate reported disease associations for 25 different genes, the authors of that pooled meta-analysis confirmed the disease associations for 8 of those genes. Interestingly, seven of these eight genes were associated with modest estimated genetic effects (OR between 1.07 and 1.76) in the pooled analyses. They concluded that there are probably many common variants in the human genome with modest but real effects on common disease risk and that studies using large samples are needed to convincingly identify such variants.

When considering the likelihood that this finding represents a true association between the SNP 11381G/C and the disease, it is important to examine the possibility of spurious effects due to chance (multiple comparisons), as well as confounding due to population stratification. Although there were seven significance tests for the primary hypothesis, these tests were not independent because these SNPs are in strong LD. It is unclear how we might make an appropriate adjustment; however, the degree of inflated type I error should be minimal. Similarly, although the observed differences in allele and genotype frequencies could be due to differences.in the genetic background of the two groups, rather than disease status (*i.e.*, population stratification), this is unlikely in this study. Genetic heterogeneity is less of a concern in Sweden than in the United States. In addition, this was a carefully designed population-based study. Almost all of the patients that met the inclusion criterion enrolled as participants in this study. Control subjects were frequency matched to cases based on residence area and age. Furthermore, the higher frequency of the risk genotypes (CC or CG) among cases diagnosed before the age of 65 years (26.3%) compared with either cases diagnosed at age 65 years or older (22.9%) or controls (19.7%) provided additional support for this association. Finally, the large number of study subjects decreases the possibility for statistical fluctuation and significantly increases our confidence in interpreting the results.

Although SNP 11381G/C is outside the NCBI-annotated 3′-UTR of *TLR4*, previous studies have not clearly defined this 3′-UTR. In fact, several expressed sequence tag clones were then identified by our BLAST search against the human expressed sequence tag database, in which we used ~2 kb of genomic sequence beyond the NCBI-annotated 3′-UTR. This computational analysis result indicated that the 3′-UTR boundary could be further downstream of the NCBI-annotated 3′ end of *TLR4*. Further support was provided by our

Table 3 *Odds ratios for prostate cancer among Toll-like receptor 4 (TLR4) single-nucleotide polymorphisms*

SNP[a] and position	Genotypes	Odd ratio[b] (95% CI)
−2026A/G	AA	1.00
	AG/GG	0.94 (0.78–1.12)
−1607T/C	CC	1.00
	TC/TT	1.08 (0.88–1.32)
3622A/C	AA	1.00
	AC/CC	0.81 (0.62–1.10)
3748A/C	AA	1.00
	AC/CC	0.96 (0.61–1.52)
7449C/A (rs2149356)	CC	1.00
	AC/AA	1.02 (0.83–1.24)
8552A/G (rs4986790), D299G	AA	1.00
	AG/GG	0.93 (0.70–1.25)
9615G/A (rs5030721)	GG	1.00
	AG/AA	1.01 (0.52–1.96)
11381G/C	GG	1.00
	CG/CC	1.26 (1.01–1.57)

[a] SNP, single-nucleotide polymorphism; CI, confidence interval.
[b] Adjusted for age and geographic regions.

experimental 3′ RACE data, which showed that the 3′-UTR of *TLR4* is 1692 bp longer than the NCBI annotation. Although SNP 11381G/C is located 78 bp beyond the 1127-bp NCBI-annotated 3′-UTR, our computational and laboratory studies show that this SNP is actually located in the center of the 2818-bp *TLR4* 3′-UTR.

The observed association of SNP 11381G/C may be due to biological impact associated with this variant, or it may indirectly reflect another unobserved causal variant of *TLR4* that is in LD with this SNP. For example, SNP 11381G/C may itself influence the stability of the mRNA species. Alternatively, other SNPs in this region that are in strong LD with SNP 11381G/C may alter an AU-rich element motif and affect mRNA stability. There are at least 20 other known SNPs in this 2818-bp 3′-UTR. A large-scale evaluation of these SNPs and functional assessments are needed to address this question. On the other hand, it is unlikely that the observed association reflects the effects of other genes in this region, as the closest known gene [deleted in bladder cancer chromosomal region (*DBCCR1*)] is 2.5 Mb 3′ from *TLR4*.

Several reported nonsynonymous changes were either not observed (E474K and Q510H) or observed only once (V310G) among our sample of 96 controls. We therefore could not assess their association with prostate cancer in our study population. Another nonsynonymous change, D299G, was relatively frequent in our population (~11% of control subjects had the variant genotypes) but was not associated with prostate cancer risk. The power calculation based on the prevalence of this SNP and the size of our study population suggested 80% power to detect an association at 5% significance level (two-sided test) if the SNP conferred at least a 1.45-fold increased risk. Multiple studies on the association between D299G and various phenotypes and diseases have been published, including the ability to recognize lipopolysaccharides, susceptibility to Gram-negative infections, premature birth, atherosclerosis, and asthma. The overall results were variable, with some reports providing positive findings (17–20) whereas others did not observe an association (21–23).

In summary, our study provided evidence of an association between a *TLR4* sequence variant and prostate cancer risk. More studies are needed to confirm or refute this finding in independent populations and to understand the mechanism by which *TLR4* sequence variants affect the expression and function of TLR4 in the signaling pathways that control innate immune response. Hopefully, our finding will also encourage additional research interest on the possible role of bacterial infection and inflammation in the development of prostate cancer.

ACKNOWLEDGMENTS

We thank all study participants in the CAPS1 study. We thank Ulrika Lund for coordinating the study at Karolinska Institute as well as all of the urologists who recruited their patients to this study and provided clinical data to the National Prostate Cancer Registry. We also thank Karin Andersson, Susan Okhravi-Lindh, Gabriella Thorén-Berglund, and Margareta Åswärd at the Regional Cancer Registries in Umeå, Uppsala, Stockholm-Gotland, and Lindköping. In addition, we thank Sören Holmgren and the personnel at the Medical Biobank in Umeå for skillfully handling the blood samples.

REFERENCES

AQ: H

1. Coussens LM, Werb Z. Inflammation and cancer. Nature (Lond) 2002;420:860–7.
2. Gardner WA Jr, Bennett BD. The prostate—overview: recent insights and speculations Monogr Pathol 1992;34:129–48.
3. De Marzo AM, Marchi VL, Epstein JI, Nelson WG. Proliferative inflammatory atrophy of the prostate: implications for prostatic carcinogenesis. Am J Pathol 1999;155:1985–92.
4. Shah R, Mucci NR, Amin A, Macoska JA, Rubin MA. Postatrophic hyperplasia of the prostate gland. Neoplastic precursor of innocent bystander? Am J Pathol 2001;158: 1767–73.
5. Carpten J, Nupponen N, Isaacs S, et al. Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. Nat Genet 2002;30:181–4.
6. Xu J, Zheng SL, Komiya A, et al. Germline mutations and sequence variants of the macrophage scavenger receptor 1 gene are associated with Prostate cancer risk. Nat Genet 2002;32:321–5.
7. Li L. Regulation of innate immunity signaling and its connection with human diseases. Curr Drug Targets Inflamm Allergy. In press 2003. AQ: I
8. Poltorak A, He X, Smirnova I, et al. Defective LPS Signaling in C3H/HeJ and C57BL/10ScCr Mice: Mutations in *Tlr4* Gene. Science (Wash DC) 1998;282:2085–8.
9. Vogel SN, Perera PY, Detore GR, et al. CD14 dependent and independent signaling pathways in murine macrophages from normal and CD14 "knockout" (CD14KO) mice stimulated with LPS or taxol. Prog Clin Biol Res 1998;397:137–46.
10. Ohashi K, Burkart V, Flohe S, Kolb H. Cutting edge: heat shock protein 60 is a putative endogenous ligand of the toll-like receptor-4 complex. J Immunol 2000;164: 558–61.
11. Arbour NC, Lorenz E, Schutte BC, et al. TLR4 mutations are associated with endotoxin hyporesponsiveness in humans. Nat Genet 2000;25:187–91.
12. Weir BS. Genetic data analysis II: methods for discrete population genetic data. Sunderland, MA: Sinauer Association, 1996.
13. Lillienfeld AM, Lillienfeld DE. Foundations of epidemiology, edition 2. New York: Oxford University Press, 1980.
14. Stephens M, Smith NJ, Donnelly PA. New statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001;68:978–89.
15. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002;70:425–34.
16. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 2003;33:177–82.
17. Schmitt C, Humeny A, Becker CM, Brune K, Pahl A. Polymorphisms of TLR4: rapid genotyping and reduced response to lipopolysaccharide of TLR4 mutant alleles. Clin Chem 2002;48:1661–7.
18. Lorenz E, Hallman M, Marttila R, Haataja R, Schwartz DA. Association between the Asp299Gly polymorphisms in the Toll-like receptor 4 and premature births in the Finnish population. Pediatr Res 2002;52:373–6.
19. Kiechl S, Lorenz E, Reindl M, et al. Toll-like receptor 4 polymorphisms and atherogenesis. N Engl J Med 2002;347:185–92.
20. Agnese DM, Calvano JE, Hahm SJ, et al. Human toll-like receptor 4 mutations but not CD14 polymorphisms are associated with an increased risk of gram-negative infections. J Infect Dis 2002;186:1522–5.
21. Read RC, Pullin J, Gregory S, Bet al. A functional polymorphism of toll-like receptor 4 is not associated with likelihood or severity of meningococcal disease. J Infect Dis 2001;184:640–2.
22. Erridge C, Stewart J, Poxton IR. Monocytes heterozygous for the Asp299Gly and Thr399Ile mutations in the Toll-like receptor 4 gene show no deficit in lipopolysaccharide signalling. J Exp Med 2003;197:1787–91.
23. Raby BA, Klimecki WT, Laprise C, et al. Polymorphisms in toll-like receptor 4 are not associated with asthma or atopy-related phenotypes. Am J Respir Crit Care Med 2002;166:1449–56.